Estadística II

Resumen

Estos son los apuntes tomados a lo largo del curso de Estadística II (4º curso de Matemáticas) en la Facultad de Ciencias de la UAM, con el profesor Jose Ramón Berrendero.

Como material adicional en las clases disponíamos de apuntes de apoyo del profesor de donde se han obtenido la mayoría de las imágenes.

Han participado en la elaboración de estos apuntes, corrigiendo erratas, completando apartados de teoría, solucionando ejercicios e incluyendo las prácticas:

- Jorge Martín
- Alberto Parramón

Agradecemos además el trabajo de todos los demás compañeros de clase que aportan sus comentarios, detección de erratas, corrección de ejercicios para aumentar la calidad de los apuntes.

Víctor de Juan

UAM - 2014/2015 2º cuatrimestre
17 de febrero de 2016 14:29

Apuntes UAM
Doble Grado Mat.Inf.

Código en Github

Índice general

[0] Documento compilado el 17 de febrero de 2016 a las 14:29

1. Introducción

1.1. Vectores aleatorios

Sea $X = (X_1, \ldots, X_p)'$ un vector aleatorio p-dimensional.

Esperanza

Definición 1.1 **Esperanza**. *Definimos la esperanza como el vector de medias, es decir:*

$$\mathbb{E}(X) = \mu = (\mu_1, \ldots, \mu_p)'$$

donde $\mu_i = \mathbb{E}(X_i)$.

Propiedades:

1. $\mathbb{E}(X + c) = \mathbb{E}(X) + c$.
2. Sea A una matriz cuadrada de dimensión $n \times p$ siendo p la dimensión de X

$$\mathbb{E}(AX) = A\mathbb{E}(X)$$

Covarianza

Definición 1.2 **Covarianza**.

$$\sigma_{i,j} = \operatorname{cov}\left(X_i, X_j\right) = \mathbb{E}\left((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$$

Dos propiedades importantes de la covarianza son:

a. $\operatorname{cov}(X, X) = \mathbb{V}(X)$

b. $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$

Al tener varias variables, ya no podemos hablar de varianzas. Definimos el correspondiente p-dimensional de la varianza.

Matriz de covarianzas

Definición 1.3 **Matriz de covarianzas**. *Llamamos* $\mathbb{V}(X) = \Sigma$ *a la matriz de covarianzas, cuya posición* (i, j) *es* $\sigma_{ij} = \operatorname{cov}\left(X_i, X_j\right)$.

Curiosidades:

- *Por la definición de covarianza, la diagonal de esta matriz es el vector p-dimensional cuya entrada* i *es la varianza de* X_i.
- *Es una matriz* ***simétrica*** *ya que* $\operatorname{cov}\left(X_i, X_j\right) = \operatorname{cov}\left(X_j, X_i\right)$

Además:

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu)(X - \mu)'] = \mathbb{E}(XX') - \mu\mu'$$

Vamos a demostrar esta última afirmación.

Demostración.

$$\mathbb{V}(X) = \mathbb{E}\left((X-\mu)(X-\mu)'\right) = \mathbb{E}(XX' - \mu X' - X\mu' + \mu\mu') =$$

$$\mathbb{E}(XX') - \mathbb{E}(\mu X') - \mathbb{E}(X\mu') + \mathbb{E}(\mu\mu') = \mathbb{E}(XX') - \mu\mathbb{E}(X') - \mu'\mathbb{E}(X) + \mu\mu' =$$

$$\mathbb{E}(XX') - \mu\mu' - \mu'\mu + \mu\mu' = \mathbb{E}(XX') - \mu\mu' = \Sigma$$

□

Propiedades: Sea X un vector aleatorio p-dimensional, A una matriz $n \times p$ y $b \in \mathbb{R}^n$

1. $\mathbb{V}(AX+b) = \mathbb{E}\left[A(X-\mu)(X-\mu)'A'\right] = A\Sigma A'$.

   Demostración.

   $$\mathbb{V}(AX+b) = \mathbb{E}\left[(AX+b-A\mu-b)(AX+b-A\mu-b)'\right] =$$

   $$= \mathbb{E}\left[(AX-A\mu)(AX-A\mu)'\right] = \mathbb{E}\left[A(X-\mu)(X-\mu)'A'\right] = A\mathbb{E}\left[(X-\mu)(X-\mu)'\right]A' =$$

   $$= A\Sigma A'$$

   □

   Con esta propiedad podemos deducir más fácilmente expresiones como $\mathbb{V}(X_1 - X_2)$ de la siguiente manera:

   $$\mathbb{V}(X_1 - X_2) = (1,-1)\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

2. Si recordamos, $\mathbb{V}(X) > 0$. La versión matricial dice Σ es semidefinida positiva.

   Demostración. Sea $a_i \in \mathbb{R}$ y $X = (X_1, \dots X_n)$ un vector aleatorio.

   $$0 \leq \mathbb{V}\left(\sum_{i=1}^{p} a_i X_i\right) = \mathbb{V}\left(a'X\right)$$

   Por la propiedad anterior, tenemos:

   $$\mathbb{V}\left(a'X\right) = a'\Sigma a$$

   , con lo que Σ tiene que ser semidefinida positiva. □

   Si Σ no es definida positiva, $\implies \exists a \in \mathbb{R}^p \nearrow a'\Sigma a = 0 \implies V(a'X) = 0 \implies \exists c \in \mathbb{R} \nearrow P(a'X = c) = 1$. Si esto se da el vector X toma valores con probabilidad 1 en un subespacio de dimensión inferior a p. En el caso de $p = 2$, las variables se situarían sobre una recta.

3. Sea $\mathbf{Y}$ un vector aleatorio $\mathbf{Y} \in \mathbb{R}^p$con matriz de covarianzas Σ, distribuido normal-multidimensionalmente. Sean $A\mathbf{Y}$ y $B\mathbf{Y}$ 2 combinaciones lineales, con $A \in \mathbb{R}^q \times \mathbb{R}^p$ y $B \in \mathbb{R}^r \times \mathbb{R}^p$. Entonces:

   $$\text{cov}\left(A\mathbf{Y}, B\mathbf{Y}\right) = A\Sigma B' = B\Sigma A'$$

Demostración.

$$\begin{pmatrix} A\mathbf{Y} \\ B\mathbf{Y} \end{pmatrix} = (A,B)'\mathbf{Y} \equiv N_{q+r}\left(\begin{pmatrix} A\mu \\ B\mu \end{pmatrix}, \begin{pmatrix} A \\ B \end{pmatrix}\Sigma\begin{pmatrix} A' & B' \end{pmatrix}\right)$$

Desarrollando la matriz de covarianzas:

$$\begin{pmatrix} A \\ B \end{pmatrix}\Sigma\begin{pmatrix} A' & B' \end{pmatrix} = \begin{pmatrix} A\Sigma A' & A\Sigma B' \\ B\Sigma A' & B\Sigma B' \end{pmatrix}$$

Y la covarianza es cualquiera de los elementos que no está en la diagonal (que como es simétrica, deberían ser iguales) □

1.2. Función característica

La función característica de un vector aleatorio X es:

$$\phi_X(t) = \mathbb{E}(e^{it'X})\ t \in \mathbb{R}^p \tag{1.1}$$

Propiedades: Lo interesante de esta función (como pasaba en el caso unidimensional) es lo siguiente:

Proposición 1.1. *Sean X e Y dos vectores aleatorios:*

$$\phi_X(t) = \phi_Y(t) \Leftrightarrow X \overset{d}{=} Y$$

Proposición 1.2 (Mecanismo de Cramér-Wold)**.**

$$a'X \overset{d}{\equiv} a'Y,\ \forall a \in \mathbb{R}^p \iff X \overset{d}{=} Y$$

Demostración.

$\Longleftarrow$ es trivial.

$\Longrightarrow$ se demuestra utilizando funciones características y tomando $t = 1$.

$$\varphi_{a'X}(t) = \varphi_{a'Y}(t),\ \forall t \in \mathbb{R}$$

$$\Longrightarrow \varphi_{a'X}(1) = \varphi_{a'Y}(1) \Leftrightarrow \mathbb{E}(e^{ia'X}) = \mathbb{E}(e^{ia'Y}) \Leftrightarrow \varphi_X(a) = \varphi_Y(a),\ \forall a \in \mathbb{R}^p$$

$$\Longrightarrow X \overset{d}{=} Y$$

□

También se cumple que:

$$X_n \xrightarrow[n\to\infty]{d} X \iff a'X_n \xrightarrow[n\to\infty]{d} a'X\ \forall a \in \mathbb{R}^n$$

1.3. Normal multivariante

Habiendo definido lo que es un vector aleatorio, vamos a definir la distribución normal multivariante, que aparecerá continuamente a lo largo del curso.

Normal p-dimensional

Definición 1.4 **Normal p-dimensional**.*El vector aleatorio X es normal p-dimensional con vector de medias μ y vector de covarianzas Σ si tiene densidad dada por:*

$$f(x) = |\Sigma|^{\frac{-1}{2}}(2\pi)^{\frac{-p}{2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \; x \in \mathbb{R}^p$$

Notación: *$X \equiv N_p(\mu, \Sigma)$ significa: X es normal p-dimensional con media μ y matriz de covarianzas Σ.*

Proposición 1.3. *Sea $\mathbf{X} \in \mathbb{R}^p$ un vector aleatorio.*

$$\mathbf{X} \sim N_p(\mu, \Sigma) \implies \forall \mathbf{a} \in \mathbb{R}^p, \mathbf{a}\mathbf{X}' \sim N_1(\mathbf{a}\mu, \mathbf{a}\Sigma\mathbf{a}')$$

Demostración.

Que la media y la varianza son esas, está calculado anteriormente en 1.

Lo de que sean una normal, Wikipedia

Corroborado por correo con José R. □

Vamos a profundizar en esta definición porque es clave, como veremos más adelante en la sección 1.4 entre otras.

Sean

$$X_1 \equiv N(0,1) \;\; X_2 = N(0,2)$$

¿El vector $\mathbf{X} = (X_1, X_2)$ cumple $\mathbf{X} \sim N_2(\mu, \Sigma)$?

Proposición 1.4. *Sean $X_i \sim N(\mu_i, \sigma_i)$*

- *X_i independientes **entonces** $\mathbf{X} = (X_1, ..., X_n) \sim N_n(\mu, \Sigma)$*

  Al ser independientes son incorreladas y por tanto la matriz de covarianzas será diagonal.
- *$\text{corr}\left(X_i, X_j\right) = 0$ y $\mathbf{X} = (X_1, ..., X_n) \sim N_n(\mu, \Sigma)$ **entonces** son independientes.*

Ejemplo: *Vamos a ver un ejemplo en dimensión 2 para ilustrar cómo reconocer si un conjunto de datos tiene una distribución normal.*

Sean $\mu = (0,0)'$ y $\Sigma = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$

Vamos a ver sus conjuntos de nivel tomando:

$$(X_1, X_2)\Sigma^{-1}(X_1, X_2)' = cte \implies \frac{x_1^2}{\lambda_1} + \frac{x_2^2}{\lambda_2} = cte$$

Dependiendo de los valores de λ_1, λ_2 tendremos casos distintos.

- *Si $\lambda_1 = \lambda_2$, entonces tendremos circunferencias.*
- *$\lambda_1 \neq \lambda_2$ entonces tendremos elipses.*

  Estas elipses tendrán como eje mayor uno de los 2 ejes, ya que las variables son independientes ($\operatorname{cov}(X_1, X_2) = 0$).

  Si por el contrario, Σ no fuera diagonal, entonces las variables no serían independientes y tendríamos una correlación entre las variables provocando que el eje mayor de la elipse fuera una recta que no corresponde con ninguno de los ejes.

Para más información consultar las transparencias de Berrendero, en las que hay un ejemplo.

1.4. Incorreladas no implica independientes

Esta sección sale de Wikipedia

Correlación

Definición 1.5 **Correlación**. *Sean X, Y dos variables aleatorias. Se define el coeficiente de correlación:*

$$\rho_{X,Y} = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Curiosidades:

- *$\rho_{X,Y} \in [-1, 1]$ Así, podemos comparar todas las correlaciones independientemente de las variables.*
- *2 variables se llaman **Variables incorreladas** si y sólo si su coeficiente de correlación es 0.*

Variables incorreladas

- *Obviamente,* $\operatorname{cov}() = 0 \implies \operatorname{corr}() = 0$

¿Incorrelación implica independencia? En general es un problema muy interesante y útil la independencia o no de variables y la correlación es algo fácil de calcular, pero **incorrelación NO implica independencia**. Esto sólo ocurre en algunos casos.

- Cuando las variables X, Y son Bernoulli, entonces incorrelación si implica independencia.[1]
- Si $\mathbf{X} = (X_1, X_2) \sim N_2(\mu, \Sigma)$ y $\operatorname{corr}(X_1, X_2) = 0$, entonces X_1 es independiente de X_2.

  La condición de $\mathbf{X} \sim N_2(\mu, \Sigma)$ es muy importante. Si sólo tuviéramos $X_i \sim N_1(\mu, \sigma)$, entonces incorrelación **no** implica independencia.

  Es por ello que este comentario se sitúa después de la definición de normal multivariante (definición 1.4)

Proposición 1.5. *Sea $\mathbf{Y}$ un vector distribuido normalmente.*

***Entonces:** un vector cualquiera $\mathbf{X}$ se distribuye normalmente si lo podemos escribir en la forma $A\mathbf{Y}$, para una matriz A.*

[1] Esto para el curso da igual, pero es interesante de saber

Demostración. Nos lo creemos del correo electrónico. □

1.5. Estandarización multivariante

Al igual que en el caso unidimensional, nos interesaría poder transformar una normal de media μ y varianza Σ en una $N(0,1)$. A continuación vamos a ver ese proceso con una normal multivariante.

Proposición 1.6 (Estandarización multivariante)**.** *Si $X \equiv N_p(\mu, \Sigma)$ y definimos $Y = \Sigma^{-1/2}(X-\mu)$, entonces $Y_1, ..., Y_p$ son i.i.d. N(0,1).*

Demostración. Sabemos por definición que:

$$f(x) = |\Sigma|^{\frac{-1}{2}}(2\pi)^{\frac{-p}{2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \; x \in \mathbb{R}^p$$

Vamos a aplicar un cambio de variable en la fórmula de la densidad:

Despejando de $Y = h(X) = \Sigma^{-1/2}(X-\mu)$, obtenemos que $\Sigma^{1/2}Y + \mu = h^{-1}(Y) = X$.

Y ahora cogemos el Jacobiano de $h^{-1}(Y) = X$ que será $\Sigma^{1/2}$ (μ es una constante e Y es la variable).

También hay que considerar la exponencial de la fórmula de la densidad, ahí hacemos el cambió de variable de:

$$e^X \text{ por } e^{h^{-1}(Y)} = e^{\Sigma^{1/2}Y+\mu}$$

Y el Jacobiano sería $e^{\Sigma^{1/2}Y}$:

Por tanto nos quedaría:

$$f(X) = f(h^{-1}(Y))\cdot\left|Jh(x)\right| = |\Sigma|^{-1/2}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}(\Sigma^{-1/2}Y+\mu-\mu)'\right)\exp\left(\Sigma^{1/2}Y\right)\Sigma^{1/2} =$$

$$= |\Sigma|^{-1/2}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}(\Sigma^{-1/2}Y)'\right)\exp\left(\Sigma^{1/2}Y\right)|\Sigma|^{1/2} =$$

$$|\Sigma|^{-1/2}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}(Y'\Sigma^{-1/2}\Sigma^{1/2}Y\right)\left|\Sigma^{1/2}\right| = (2\pi)^{-p/2}\exp\left(-\frac{1}{2}(Y'Y)\right)$$

□

Vamos a ver un ejemplo para profundizar en la distribución.

Ejemplo: *Definimos el siguiente vector aleatorio: $X = (X_1, X_2, X_3)' \equiv N_3(\mu, \Sigma)$ con:*

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 7/2 & 1/2 & -1 \\ 1/2 & 1/2 & 0 \\ -1 & 0 & 1/2 \end{pmatrix}$$

a) *Calcula las distribuciones marginales $X_i \equiv N(\mathbb{E}(X_i), \mathbb{V}(X_i))$:*

$X_1 \equiv N(0, 7/2)$

$X_2 \equiv N(0, 1/2)$

$X_3 \equiv N(0, 1/2)$

Para calcular estos valores solo hace falta mirar los datos que nos da el problema, el vector de medias μ y la matriz de covarianzas Σ:

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \mathbb{V}(X_2) & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \mathbb{V}(X_3) \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \mathbb{E}(X_3) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

b) *Calcula la distribución del vector $(X_1, X_2)'$:*

Este vector sigue una distribución normal que puede obtener de las matriz Σ y el vector de medias μ:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \equiv N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 7/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \right]$$

c) *¿Son X_2 y X_3 independientes?*

Sí son independientes ya que la covarianza entre ambas variables es 0. La covarianza entre X_2 y X_3 es el elemento de la fila 3 y la columna 2 de la matriz de covarianzas Σ, (que al ser Σ simétrica coincide con el elemento de la fila 2 y la columna 3).

d) *¿Es X_3 independiente del vector $(X_1, X_2)'$?*

No son independientes ya que el vector de covarianzas entre ambas variables no es 0. Como en el caso anterior, tomamos como el elemento que ocupa la fila 3 y las columnas 1 y 2, es decir, el vector $(-1, 0)$, que al no ser idénticamente nulo, concluimos que X_3 no es independiente del vector (X_1, X_2)

e) *Calcula la distribución de la variable aleatoria $(2X_1 - X_2 + 3X_3)$.*

Procedemos de la siguiente manera:

$$(2X_1 - X_2 + 3X_3) = (2, -1, 3) \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \equiv N(0,)$$

1.5.1. Consecuencias de la estandarización

- La función característica de una v.a. normal $N_p(\mu, \Sigma)$ viene dada por:

$$\varphi_X(t) = \exp\{it'\mu - \frac{1}{2}t'\Sigma t\}$$

Demostración. Tenemos $X \sim N_p(\mu, \Sigma)$ con $Y \sim N_p(0, I)$, y con $X = \Sigma^{\frac{1}{2}} Y + \mu$. Aplicando 1.1 obtenemos:

$$\varphi_X(t) = \mathbb{E}\left[e^{it'(\Sigma^{\frac{1}{2}}Y+\mu)}\right] = e^{it'\mu} \cdot \mathbb{E}\left[e^{it'\Sigma^{\frac{1}{2}}Y}\right] \underbrace{=}_{u=\Sigma^{1/2}t} e^{it'\mu} \cdot \mathbb{E}\left[e^{i(u_1Y_1+\ldots+u_pY_p)}\right]$$

Teniendo en cuenta que $Y_1, \ldots, Y_p \overset{\text{iid}}{\sim} N(0,1)$, y que la función característica de una normal estándar unidimensional ($X \sim N(0,1)$) es $\varphi_X(t) = e^{-\frac{t}{2}}$ tenemos:

$$e^{it'\mu} \cdot \prod_{j=1}^{p} \varphi_{Y_j}(u_j) = e^{it'\mu} \cdot \prod_{j=1}^{p} e^{-\frac{u_j^2}{2}} =$$

$$= \exp\{it' - \frac{1}{2}u'u\} = \exp\{it'\mu - \frac{1}{2}t'\Sigma t\}$$

□

- La distribución de $(X-\mu)'\Sigma^{-1}(X-\mu)$ es χ^2_p.

Demostración. Tenemos $X \sim N_p(\mu, \Sigma)$ y $X = \Sigma^{\frac{1}{2}} Y + \mu$, por tanto:

$$(X-\mu)'\Sigma^{-1}(X-\mu) = Y'\Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2}Y = Y'Y = \sum_{i=1}^{p} Y_i^2 \sim \chi^2_p$$

Pues sabemos que $Y_1, \ldots, Y_p \overset{\text{iid}}{\sim} N(0,1)$ y que χ^2_p se construye como la suma de p normales estándar independientes. □

1.6. Distancia de Mahalanobis

La distancia de Mahalanobis tiene como objetivo introducir la variabilidad y correlaciones de las variables a la hora de ver cuánto dista un dato de otro sabiendo que ambos provienen de cierta distribución.

Distancia de Mahalanobis

Definición 1.6 **Distancia de Mahalanobis**. *La distancia de Mahalanobis de un vector aleatorio a su vector de medias se define como:*

$$d_M(X, \mu) = \sqrt{(X-\mu)'\Sigma^{-1}(X-\mu)}$$

Tiene las siguientes **propiedades**:

1. d_M coincide con la distancia euclídea entre los datos estandarizados de forma multivariante
2. d_M es adimensional
3. d_M tiene en cuenta las diferentes variabilidades (varianzas) de las variables
4. d_M tiene en cuenta las correlaciones entre las variables
5. d_M bajo normalidad, su cuadrado se distribuye como una χ^2_p

Vamos a comprobar la primera y última propiedad:

Demostración.

- d_M coincide con la distancia euclídea entre los datos estandarizados de forma multivariante:

  Tenemos $X \sim N(\mu = 0, \Sigma = I)$ y por tanto:

$$d_M(X, \mu = 0) = \sqrt{X'IX} = \|X\|$$

- d_M bajo normalidad, su cuadrado se distribuye como una χ^2_p:

  Tomando $Y = \Sigma^{-1/2}(X - \mu)$ se tiene $Y \sim N(0, I)$, y por tanto:

$$d^2_M(X, \mu) = (X - \mu)'\Sigma^{-1}(X - \mu) = \left[\Sigma^{-1/2}(X - \mu)\right]' \left[\Sigma^{-1/2}(X - \mu)\right] =$$

$$= Y'Y = \sum_{j=1}^{p} Y_j^2 \sim \chi^2_p$$

□

Observación: Si al aplicar la d_M a cada uno de los datos de la observación, vemos que el histograma no corresponde con la función de densidad de una χ^2_p, se puede deducir que $X \overset{d}{\nsim} N_p(\mu, \Sigma)$.

1.7. Transformaciones afines de vectores normales

Proposición 1.7. *Si $X \equiv N_p(\mu, \Sigma)$, A es matriz $q \times p$ y $b \in \mathbb{R}^q$, entonces:*

$$AX + b \equiv N_p(A\mu + b, A\Sigma A')$$

Demostración.

$$\begin{aligned}\varphi_{AX+b}(t) &= \mathbb{E}(e^{it'(AX+b)}) = e^{it'b} \cdot \mathbb{E}(e^{it'AX}) \\ &= e^{it'b} \cdot \varphi_X(A't) = e^{it'b} \cdot \exp\{it'\mu - \frac{1}{2}t'A\Sigma A't\} \\ &= \exp\{it'(A\mu + b) - \frac{1}{2}t'A\Sigma A't\}\end{aligned}$$

Que corresponde precisamente con la función característica de una normal $N_p(A\mu + b, A\Sigma A')$. □

Corolario 1.8. *Si $X = (X_1|X_2)$, con $X_1 \in \mathbb{R}^q$ y $X_2 \in \mathbb{R}^{p-q}$,y consideramos las particiones correspondientes de μ y Σ:*

$$\mu = (\mu_1|\mu_2), \Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

Entonces $X_1 \equiv N(\mu_1, \Sigma_{11})$.

Proposición 1.9. *Dichas X_1 y X_2 son independientes* $\Leftrightarrow \Sigma_{12} = 0$.

Demostración. *Hecho por Jorge. Se aceptan correcciones.*

Se demuestra la implicación hacia la izquierda (la otra es equivalente pero empezando por el final).

$$f(X) = |\Sigma|^{-1/2}\,(2\pi)^{-p/2}\exp\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\} =$$

$$= |\Sigma|^{-1/2}\,(2\pi)^{-p/2}\exp\left\{-\frac{1}{2}(X_1-\mu_1, X_2-\mu_2)\begin{bmatrix}\Sigma_{11}^{-1} & \Sigma_{12}^{-1}\\ \Sigma_{21}^{-1} & \Sigma_{22}^{-1}\end{bmatrix}\begin{pmatrix}X_1-\mu_1\\ X_2-\mu_2\end{pmatrix}\right\} =$$

$$= |\Sigma|^{-1/2}\,(2\pi)^{-p/2}\exp\{-\frac{1}{2}[(X_1-\mu_1)'\Sigma_{11}^{-1}(X_1-\mu_1) + (X_2-\mu_2)'\Sigma_{21}^{-1}(X_1-\mu_1)$$
$$+ (X_1-\mu_1)'\Sigma_{12}^{-1}(X_2-\mu_2) + (X_2-\mu_2)'\Sigma_{22}^{-1}(X_2-\mu_2)]\}$$

Con $\Sigma_{12} = \Sigma_{21} = 0$ se tiene:

$$f(X) = |\Sigma_{11}|^{-1/2}\,(2\pi)^{-q/2}\exp\left\{-\frac{1}{2}(X_1-\mu_1)\Sigma_{11}^{-1}(X_1-\mu_1)\right\}\cdot$$
$$\cdot\,|\Sigma_{22}|^{-1/2}\,(2\pi)^{-(p-q)/2}\exp\left\{-\frac{1}{2}(X_2-\mu_2)\Sigma_{22}^{-1}(X_2-\mu_2)\right\} =$$
$$= f(X_1)\cdot f(X_2)$$

Por tanto tendremos que X_1 y X_2 son independientes. □

Observación:

- Si dos v.a. X e Y tienen distribución normal y además $Cov(X,Y) = 0$, esto no implica que X e Y sean independientes. Para que esto sea cierto es necesario que X e Y sean conjuntamente normales.
- Si dos v.a. X e Y tienen distribución normal y $a, b \in R$, la combinación lineal $aX + bY$ no tiene necesariamente distribución normal. Para que esto sea cierto es necesario que X e Y sean independientes.

Demostración. *Hecho por Jorge. Se aceptan correcciones.*

Con independencia se tiene:

$$\varphi_{aX+bY}(t) = \mathbb{E}[e^{it(aX+bY)}] = \mathbb{E}[e^{itaX}]\mathbb{E}[e^{itbY}] =$$
$$= \varphi_X(ta)\cdot\varphi_Y(tb) = \exp\{ita\mu_X - \frac{1}{2}\sigma_X^2t^2a^2\}\cdot\exp\{itb\mu_Y - \frac{1}{2}\sigma_Y^2t^2b^2\} =$$
$$= \exp\{it(a\mu_X + b\mu_Y) - \frac{1}{2}(\sigma_X^2a^2 + \sigma_Y^2b^2)t^2\}$$

Que coincide con la función característica de una normal:

$$N(a\mu_X + b\mu_Y, \sigma_X^2a^2 + \sigma_Y^2b^2)$$

□

- Aunque todas las marginales de un vector aleatorio p-dimensional X tengan distribución normal, esto no implica que X tenga distribución normal p-dimensional.Para que esto sea cierto es necesario que X e Y sean independientes. Se demuestra en el ejercicio 4 de la hoja 1 1.4.

1.8. Distribuciones condicionadas

Proposición 1.10. *Sea $X = (X_1, X_2)$ con $X_1 \in \mathbb{R}^p$ y $X_2 \in \mathbb{R}^{p-q}$.*

$$\mu = (\mu_1, \mu_2)$$
$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

entonces: *$X_2|X_1 \sim N_{p-q}(\mu_{2.1}, \Sigma_{2.1})$, donde*

$$\mu_{2.1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1) = \mathbb{E}\left(X_2|X_1\right) \tag{1.2}$$

$$\Sigma_{2.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \mathbb{V}\left(X_2|X_1\right) \tag{1.3}$$

Demostración. Definimos $X_{2.1} = X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1$.

$$\begin{pmatrix} X_1 \\ X_{2.1} \end{pmatrix} = \left(\begin{array}{c|c} I & 0 \\ \hline -\Sigma_{21}\Sigma_{11}^{-1} & I \end{array}\right)$$

Como es una combinación lineal de $(X_1, X_2)'$, entonces $X_{2.1}$ es normal multivariante.

Vamos a calcular la media y la matriz de covarianzas de $X_{2.1}$

$$X_{2-1} = N\left(\mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \left(\begin{array}{c|c} \Sigma_{11} & 0 \\ \hline 0 & \Sigma_{2.1} \end{array}\right)\right)$$

Donde las covarianzas se calculan: $A\Sigma A'$, siendo A la matriz de la combinación lineal, es decir:

$$A = \left(\begin{array}{c|c} I & 0 \\ \hline -\Sigma_{21}\Sigma_{11}^{-1} & I \end{array}\right)$$

Conclusiones:

- X_1 es independiente de $X_{2.1}$

- $X_{2.1}$ es normal, con media y varianza calculadas anteriormente.

  $X_{2.1}|X_1$, al ser independientes, también se distribuye normalmente, con los mismos parámetros.

- Dado X_1, los vectores $X_{2.1}$ y X_2 difieren en el vector constante $\Sigma_{21}\Sigma_{11}^{-1}X_1 \implies X_2|X_1 = N(\mu_{2.1}, \Sigma_{2.1})$

□

Ejemplo: *Vamos a considerar X_1, X_2 como escalares, para entender la proposición. Este ejemplo le surgió a un investigador que quería predecir la estatura de los hijos en función de la de los padres (que no padres y madres, sólo padres).*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \equiv N_2 \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \right)$$

Definimos

$$\overline{Y} = \mathbb{E}\left(Y|X\right) = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x)$$

. La esperanza de la altura del hijo condicionada a la altura del padre será la media de las alturas de los hijos corregida por un factor en el que influye la diferencia de altura del padre con respecto a su media. Es de esperar que si Yao Ming tiene un hijo, sea más alto que la media.

El factor de corrección $\frac{\sigma_{xy}}{\sigma_x^2}$ es importante y no me he enterado bien de dónde sale.

Ahora vamos a calcular $\mathbb{V}\left(Y|X\right) = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2} = \sigma_y^2\left(1 - \rho^2\right)$ donde $\rho = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2}$, el coeficiente de correlación.

Ha dicho algo así como La única relación que puede existir entre 2 variables normales es una relación lineal.

Este coeficiente de correlación aparece también en la expresión de la esperanza. Vamos a verlo:

$$\overline{Y} = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x) \iff \frac{\overline{Y} - \mu_y}{\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\frac{x - \mu_x}{\sigma_x}$$

Es decir:

$$\frac{\overline{Y} - \mu_y}{\sigma_y} = \rho\frac{x - \mu_x}{\sigma_x}$$

Aplicado a la estatura de los hijos respecto de los padres, se interpreta como: "Si un padre es muy alto, su hijo será alto pero no destacará tanto como el padre". Este fenómeno lo definió como ***Regresión a la mediocridad****.*

Regresión a la mediocridad

Homocedasticidad *Definición 1.7* **Homocedasticidad**. *Sea $X = (X_1, X_2)$ con $X_1 \in \mathbb{R}^p$ y $X_2 \in \mathbb{R}^{p-q}$. Entonces son vectores* ***homocedásticos*** $\iff$ *$\Sigma_{2.1}$ es constante.*

Ya veremos más adelante este concepto con mayor detalle.

Ejemplo: *Ahora vamos a ver un par de ejemplos numéricos:*

Sea

$$\left(X,Y\right) \equiv N_2\left(\left(0,0\right), \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}\right)$$

Distribución $Y|X$: *Utilizando las fórmulas definidas en 1.10 para X_i unidimensionales:*

$$\mathbb{E}\left(Y|X\right) = \mu_{2-1} = 0 + 3 \cdot \frac{1}{10}(X-0) = \frac{3}{10}x$$
$$\mathbb{V}\left(Y|X\right) = \Sigma_{2.1} = 1 - \frac{3}{10} \cdot 3 = \frac{1}{10}$$

Distribución $X|Y$:

$$E(X|Y) = 3y$$
$$V(X|Y) = 1$$

Ambas son normales unidimensionales ya que (X,Y) es normal multivariante.

Ejemplo: *Sea*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

Queremos calcular la distribución de $(X+Y)|(X-Y)=1$

Para ello, definimos 2 variables, $Z_1 = X+Y$ y $Z_2 = X-Y$, con lo que ahora tenemos que calcular $Z_2|Z_1 = 1$

Lo primero es hallar la relación matricial entre X, Y y Z_i

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X+Y \\ X-Y \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}$$

¿Cuáles son la esperanza y la matriz de covarianzas de el vector aleatorio (Z_1, Z_2)? Para ello necesitamos la matriz de la combinación lineal que ya tenemos:

$$\mu_Z = A \cdot \mu_{xy} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

$$\Sigma_Z = A\Sigma_{xy}A' = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}$$

Ahora ya podemos calcular la distribución como en el ejemplo anterior:

$$\mathbb{E}\left(Z_1|Z_2=1\right)=2+1\cdot\frac{1}{3}(1-0)=\frac{7}{3}$$
$$\mathbb{V}\left(Z_1|Z_2=1\right)=7-1\cdot\frac{1}{3}\cdot 1=\frac{20}{3}$$

En este caso, al ser homocedásticas (1.7) entonces $\mathbb{V}\left(Z_1|Z_2=1\right)=\mathbb{V}\left(Z_1|Z_2=n\right)\forall n\in\mathbb{N}$

1.9. Formas cuadráticas bajo normalidad

Proposición 1.11. *Bajo las hipótesis:*

a. B *una matriz simétrica e idempontente*

b. $Y\sim N_2(\mu,\sigma^2 I_n)$

c. $\mu' B\mu = 0$

se tiene que:

$$\frac{Y'BY}{\sigma^2}\equiv\chi^2_p$$

con $p=Rg(B)$.

Observación:

- La única matriz idempontente de rango completo es I_n.
- $\lambda=0,1\ \forall\lambda$ autovalor de B.

Demostración.

$$\left.\begin{array}{c} Bu=\lambda u \\ Bu=B^2u=\lambda Bu=\lambda^2 u \end{array}\right\}\lambda u=\lambda^2 u\implies\lambda=0,1$$

□

- Este último hecho permite calcular los grados de libertad de la distribución más fácilmente, ya que $p=Rg(B)=tr(B)=\#\{i\ /\ \lambda_i=1\}$

Demostración. Puesto que B es simétrica e idempotente tenemos que se puede expresar como $B=CDC'$ donde C es una matriz formada por autovectores ortonormales y D es la matriz diagonal con los autovalores de B (que como ya hemos visto serán exclusivamente $\lambda=0,1$).

$$B=CDC'=\underset{n\times p}{(C_1,C_2)}\left(\begin{array}{c|c} I_p & 0 \\ \hline 0 & 0 \end{array}\right)\left(\begin{array}{c} C_1 \\ \hline C_2 \end{array}\right)=C_1C_1'$$

Puesto que C_1 está formado por autovectores ortogonales se tiene que:

$$C_1'C_1=I_p$$

$$Y'BY = Y'C_1 \underbrace{C_1'Y}_{Z} = Z'Z = \sum_{i=1}^{p} z_i^2$$

$$Z \sim N_p(C_1'\mu, \sigma^2 I_p)$$

Pero se puede ver que la media es nula haciendo uso de la tercera hipótesis:

$$\left\|C_1'\mu\right\|^2 = \mu' C_1 C_1'\mu = \mu' B\mu = 0 \implies C_1'\mu = 0$$

Así ya solo falta dividir por σ^2 para llegar a que $z_1, \ldots, z_p \overset{\text{iid}}{\sim} N(0,1)$ y por tanto:

$$\frac{Y'BY}{\sigma^2} \equiv \chi_p^2$$

□

Proposición 1.12 (Formas cuadráticas bajo normalidad). *Sea* $\mathbf{Y} \equiv N_n(\mu, \sigma^2 I_n)$ *y sean* $A_{p\times n}, B_{q\times n}, C_{n\times n}, D_{n\times n}$ *con* C, D *simétricas e idempotentes.*

Entonces:

- AY *y* BY *son independientes* $\iff AB' = 0$
- $Y'CY$ *e* $Y'DY$ *son independientes* $\iff CD = 0$
- AY *e* $Y'CY$ *son independientes* $\iff AC = 0$

Demostración.

- *La primera afirmación se demuestra realizando una transformación lineal sobre* Y*:*

$$\begin{pmatrix} A \\ B \end{pmatrix} Y \sim N\left(\begin{pmatrix} A \\ B \end{pmatrix}\mu, \sigma^2 \cdot \left(\begin{array}{c|c} AA' & AB' \\ \hline BA' & BB' \end{array}\right)\right)$$

  Por tanto tendremos independencia $\iff AB' = 0$.

- *La segunda afirmación se deduce como consecuencia de la primera pues:*

$$Y'CY = \|CY\|^2 = Y'\underbrace{C'C}_{C}Y$$

$$Y'DY = \|DY\|^2$$

- *La tercera afirmación es consecuencia directa de las dos anteriores.*

□

Lema 1.13 (Lema de Fisher). *Sean* $Y_1, .., Y_n \overset{iid}{\sim} N(\mu, \sigma)$*. Vamos a considerar el vector cuyas marginales son estas* $Y \equiv (Y_1, ..., Y_n) = N(\mu 1_n', \sigma^2 I_n)$[2]

[2] $1_n = (1, 1, ..., 1)$ n veces.

Entonces: $\overline{Y}, S^2 = \frac{\sum(Y_i-\overline{Y})^2}{n-1}$ *son independientes. Además,*

$$\frac{(n-1)S^2}{\sigma^2} \equiv \chi^2_{n-1}$$

Demostración. Vamos a tomar $\overline{Y} = \frac{1}{n}1_n Y$:

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2 = \left(\sum_{i=1}^{n}Y_i^2\right) - n\overline{Y}^2 = Y'Y - Y'\underbrace{\frac{1}{n}1_n'1_n}_{M}Y =$$

Donde M es una matriz $n \times n$:

$$M = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

$$= Y'Y - Y'MY = Y'(I-M)Y$$

Tenemos que M es una matriz simétrica e idempotente cuyas columnas generan el subespacio sobre el que proyectamos (en este caso el subespacio $V \equiv< (1,\ldots,1) >$). Y por tanto $I-M$ también será matriz simétrica e idempotente.

De modo que para poder aplicar la proposición 1.12 debemos ver si $(\mu 1_n)(I-M)(\mu 1_n)' = 0$:

$$\mu^2 1_n(I-M)1_n' = \mu^2 1_n(1_n' - 1_n') = 0$$

Lo siguiente es ver el rango de $I-M$:

$$rg(I-M) = tr(I-M) = tr(I) - tr(M) = n-1$$

Ahora ya estamos en condiciones de aplicar la proposición 1.12 para obtener que:

$$\frac{\sum(Y_i-\overline{Y})^2}{\sigma^2} = \frac{Y'(I-M)Y}{\sigma^2} \equiv \chi^2_{n-1}$$

□

Teorema 1.14 (Teorema central del límite Multivariante)**.** *Sean $X_1, ..., X_n$ vectores aleatorias independientes e idénticamente distribuidas (vec.a.i.i.d.) con $X_i \sim N(\mu, \Sigma)$, con Σ definida positiva.*

Entonces:

$$\sqrt{n}\Sigma^{\frac{-1}{2}}(\overline{X_n} - \mu) \xrightarrow[n\to\infty]{d} N(0, I) \iff \sqrt{n}(\overline{X_n} - \mu) \xrightarrow[n\to\infty]{d} N_p(0, \Sigma)$$

La velocidad a la que $\overline{X_n}$ converge a μ es del orden de $\frac{1}{\sqrt{n}}$

Demostración. Lamentándolo mucho, la prueba será ignorada por el momento. □

Capítulo I

Contrastes no paramétricos

1. Introducción

Hipótesis no paramétrica

Definición 1.1 **Hipótesis no paramétrica**. *Hipótesis que no se formula en términos de un número finito de parámetros.*

En este capítulo vamos a ver 3 tipos de contrastes (con diferentes test para algunos contrastes)

1. **Bondad de ajuste:** A partir de una muestra $X_1, ..., X_n \equiv F$ de variables aleatorias independientes idénticamente distribuidas, contrastar:
   - $H_0 : F = F_0$ donde F_0 es una distribución prefijada.
   - $H_0 : F \in \{F_\theta : \theta \in H\}$, donde H es el espacio paramétrico.
2. **Homogeneidad:** Dados $X_1, ..., X_n \equiv F$ y $Y_1, ..., Y_n \equiv G$ de variables aleatorias independientes idénticamente distribuidas. Contrastar $H_0 : F = G$.
3. **Independencia:** Dada $(X_1, Y_1), ..., (X_n, Y_n) \equiv F$ de variables aleatorias independientes idénticamente distribuidas. Contrastar $H_0 : X$ e Y son independientes.

2. Contrastes de bondad de ajuste

2.1. Contraste χ^2 de bondad de ajuste

Consideramos una distribución totalmente especificada bajo $H_0 : X_1, ..., X_n \equiv F$ de variables i.i.d.

$H_0 : F = F_0$ es la hipótesis nula y queremos ver que F, que es la distribución obtenida con los datos verdaderos (obtenidos empíricamente) es igual a F_0 que es la distribución teórica.

Vamos a definir los pasos que tenemos que seguir para comprobar si H_0 es cierta:

1. Se definen k clases $A_1, ..., A_k$. En el caso del dado, los valores de cada cara.
2. Se calculan las frecuencias observadas de datos en cada clase.

$$O_i = \#\{j \ / \ X_j \in A_i\}$$
$$O_i \sim Bin\left(n, p_i = p_{H_0}(A_i)\right) \text{ es una variable aleatoria}$$

3. Se calculan las frecuencias esperadas si H_0 fuera cierta:

$$\mathbb{E}(F)_i = \mathbb{E}(F)_{H_0} = np_i \text{ al ser } O_i \text{ una binomial}$$

4. Se comparan O_i con E_i mediante el estadístico de Pearson, para comprobar si lo observado se parece a lo esperado.

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Más adelante justificaremos porqué este estadístico es el utilizado. Además, el estadístico puede calcularse de otra manera:

$$T = \sum \frac{O_i^2}{E_i} - n$$

5. Se rechaza H_0 en la región crítica $R = \{T > c\}$, donde c depende del nivel de significación α.

c se obtiene consultando en las tablas para $\alpha = P_{H_0}(T > c)$

Pero se nos presenta el siguiente problema, ¿Cuál es la distribución de T bajo H_0?

Teorema 2.1 (Estadístico de Pearson). *El estadístico de Pearson es:*

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

Además, bajo H_0 su distribución:

$$T \xrightarrow[n\to\infty]{d} \chi^2_{k-1}$$

Demostración. Para otro momento □

Ejemplo: *Tiramos un dado 100 veces y obtenemos:*

Resultados	*1*	*2*	*3*	*4*	*5*	*6*
Frecuencia	*10*	*20*	*20*	*10*	*15*	*25*

Y consideramos $H_0 : p_i = \frac{1}{6}\ \forall i = 1, ..., 6$. Es decir que el dado no está trucado y cada cara tiene la misma probabilidad (p_i) de salir.

¿Es cierta la hipótesis, con un nivel de confianza/significación del 95 %?

Las clases son cada uno de los posibles resultados y las frecuencias observadas se encuentran en la tabla.

Vamos a calcular el estadístico T

$$T = \sum_{i=1}^{6} \frac{O_i^2}{E_i} - n = \frac{6}{100}(\sum O_1^2) - 100 = ... = 11$$

Ahora, consultando las tablas buscamos el valor $\chi^2_{5;0.05} = 11.07 > T = 11$*, entonces no estamos en la región crítica, por lo que no podemos rechazar la hipótesis.*

Al ser valores muy próximos, observamos que el p-valor[1] *del contraste tendrá que ser algo mayor que* 0.05*.*

2.1.1. Hipótesis nula compuesta

Vamos a estudiar el siguiente problema. Sea $X_1, ..., X_n \sim iidF$ y una hipótesis compuesta: $H_0 : F \in \{F_\theta, \theta \in \Omega \subset \mathbb{R}^n\}$. Esta hipótesis compuesta puede ser "los datos se distribuyen normalmente, con media y varianza desconocidas"

Los pasos a seguir son:

1. Definir las clases $A_1, ..., A_k$
2. Calcular las frecuencias observadas $O_1, ..., O_k$
3. Estimamos θ por el método de máxima verosimilitud. Sea $\overline{\theta}$ el e.m.v.

   Pero para calcular las frecuencias esperadas, no tenemos una única normal. La idea intuitiva sería: hay unos parámetros que son los que mejor ajustan la distribución. ¿Cuál es la que mejor ajusta? La que tenga los parámetros estimados.
4. Ahora ya podemos calcular las frecuencias esperadas:

   $E_i = n\overline{p_i}$ donde $\overline{p_i} = P_\theta(A_i)$, con $i = 1, ..., k$
5. Ya podemos calcular el estadístico de Pearson:

$$T = \sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i}$$

   ¿Qué distribución tiene este estadístico? Antes hemos visto que es una χ^2_{k-1} cuando se dan unas ciertas condiciones.

   En este caso, es de esperar que T tienda a tomar valores menores que en el caso simple.

   Además, al estimar r parámetros (las r componentes del vector θ)) se introducen r nuevas restricciones sobre el vector $(O_1, ..., O_r)$.

   Se puede probar, [2] que:

$$\sum_{i=1}^{k} \frac{(O_i - \hat{E}_i)^2}{\hat{E}_i} \xrightarrow[n\to\infty]{d} \chi^2_{k-1-r}$$

[1] el menor nivel de confianza para poder rechazar

[2] bajo ciertas condiciones de regularidad que las distribuciones que conocemos cumplen y que son demasiado complicadas de enunciar

6. Se rechaza H_0 en la región crítica

$$R = \{T > \chi^2_{k-1-r}\}$$

Ejemplo:

Los bombardeos de Londres *Los alemanes bombardeaban mucho a Londres durante la guerra mundial, y los ingleses querían saber si los alemanes podían dirigir los misiles, o los impactos eran aleatorios.*

Para ello, alguien hizo el estudio estadístico, para contrastar la hipótesis "los impactos son aleatorios".

Los impactos deberían seguir una Poisson [3]*. Para ello, dividió Londres en* $n = 576$ *cuadrados, cada uno de ellos será la variable* X_i *que debería seguir una Poisson.*

La idea del contraste es: la Poisson que más se puede parecer es la que tenga de media el e.m.v. Si esa Poisson no se parece, entonces ninguna Poisson se puede parecer.

Clases *Los valores que toma la Poisson (recordamos que son número naturales). En este caso sólo se han definido 5, ya que la última es* > 4 [4]

Obs $O_i = \{\#j : X_j = i\}$*, por ejemplo, las frecuencias observadas de la clase* 0*, es decir,* $O_0 = 229$*, donde ese número es el número de las* n *regiones de Londres en donde no cayó ningún misil.*

e.m.v. *El e.m.v. de una Poisson es la media muestral, con lo que* $\hat{\lambda} = \overline{x} = 0.9323$

Esp *Calculamos las frecuencias esperadas utilizando el parámetro estimado*

$$\hat{E}_i = n\hat{p}_i = 576 \cdot e^{-0.9323}\frac{(0.9323)^i}{i!}$$

T *El estadístico* χ^2 *de Pearson es:*

$$T = \sum_{i=1}^{k} \frac{(O_i - \overline{E_i})^2}{\overline{E_i}} = 1.01$$

Desde esta información, ya podemos hacernos una idea de si vamos a poder rechazar. ¿Por que? Al estimar un único parámetro $T \overset{iid}{\sim} \chi^2_{5-1-1}$*, y además* $\mathbb{E}(\chi^2_k) = k$*, con lo que debería habernos salido* $T = 3$*. Al ser un vector bastante normal, podemos ver que tiene muy poca pinta de que vayamos a poder rechazar la hipótesis nula. Al* T *estar por debajo de* 3 *no estaremos en la región crítica.*

Observación*: Este ejemplo se encuentra también en las transparencias, donde podemos ver los valores y algunas gráficas explicativas.*

[3] ya que es el límite de una binomial, en la que consideramos los impactos como éxitos

[4] Se recomienda no definir más de 5 clases, para que la estimación no pierda demasiada información.

2.1.2. Contrastes con R

Hipótesis simple Con el siguiente código de R, podemos hacer el contraste de bondad de ajuste de una χ^2 fácilmente.

```
obs = c(10,20,20,10,15,25)
ls.str(chisq.test(obs))
```

Si a *chisq.test* no le damos más argumentos, supondrá hipótesis simple con equiprobabilidad de p. Podríamos darle otro argumento, y hacer lo siguiente para el ejemplo de los misiles:

```
res = c(seq(0,4),7)
obs = c(229,211,93,35,7,1)
n = sum(obs)
lambda = sum(res*obs)/n
prob = dpois(res,lambda)
esp = n*prob
# Se agrupan las dos ultimas clases:
obs = c(obs[1:4],sum(obs[5:6]))
prob = c(p[1:4],1-sum(p[1:4]))
esp = c(esp[1:4],n-sum(esp[1:4]))
# Codigo para el grafico de barras:
matriz = rbind(p,obs/n)
rownames(matriz) = c('Frecuencias','Poisson')
barplot(matriz,beside=TRUE,names.arg=c(0:4),legend.text=TRUE,
col=c('lightgreen','orange'))
# Test chi 2
t = chisq.test(obs,p=prob)$statistic
pvalor = 1 - pchisq(t,3)
```

2.2. Kolmogorov-Smirnov

$X_1, ..., X_n \overset{iid}{\sim} F$ y $H_0 : F = F_0$, con F_0 totalmente especificada.

Se define la función de distribución empírica correspondiente a X_i como

$$F_n(x) = \frac{1}{n}\#\{i : X_i \leq x\}$$

Ejemplo: $X_1 = 1, X_2 = 6, X_3 = 4$

Estadístico de orden

Definición 2.1 ***Estadístico de orden.***

Los estadísticos de orden, consisten en ordenar la muestra y se escriben: $X_{(i)}$.

Esto significa el X_j *que ocupa la posición* $i - esima$ *cuando ordenamos de menor a mayor. Es decir,* $X_{(1)} = \text{mín}\{X_j\}$ *y* $X_{(n)} = \text{máx}\{X_j\}$ *Es una función de distribución constante a trozos con saltos de magnitud* $\frac{1}{n}$ *en cada valor muestral* X_i *(si no hay empates). En este caso, la función de distribución es* $f(a) = \frac{1}{3}, a = \{1, 4, 6\}$

F_n *es un estimador de la verdadera distribución de* F

$$F(x) = P(X \leq x)$$

Observación:

$$\#\{i : X_i \leq x\}$$

¿Qué distribución tiene esta variable aleatoria? Es una binomial con parámetros: $B(n, F(x))$

$$\mathbb{E}\left(F_n(x)\right) = \frac{1}{n} nF(x) = F(x)$$

Vemos que la estimación es insesgada. Además,

$$\mathbb{V}\left(F_n(x)\right) = \frac{1}{n^2} nF(x)(1 - F(x)) \xrightarrow[n\to\infty]{} 0$$

Estimador consistente

Con lo que, $F_n \xrightarrow[n\to\infty]{P} F(x)$ y decimos que es un **Estimador consistente**[5]

¿Esto a qué viene? A entender la función de distribución empírica.

Además de la convergencia en probabilidad tenemos un resultado más potente:

Teorema 2.2 (Teorema Glivenco-Cantelli)**.**

$$||F_n - F||_\alpha = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{c.s.} 0$$

Es decir, la función de distribución empírica converge en distribución a la real y además esa convergencia es uniforme.

Demostración. La demostración se sale del temario de este curso. □

¿Cómo se hace realmente el contraste? Si H_0 fuese cierta, se espera que $D_n = ||F_n - F_0||$ sea "pequeño". La idea es rechazar en la región crítica $R = \{D_n > c\}$ para un valor c tal que $P_{H_0}(D_n > c) = \alpha$, siendo α un nivel de significación predeterminado.

Para poder calcular $P_{H_0}(D_n > c)$ necesitamos cálculos algo chungos, pero hay un resultado que nos puede ayudar mucho a resolver este problema.

Lema 2.3. *La distribución bajo H_0 de D_n es la misma para cualquier distribución continua F_0*

Por tanto el valor de c en la región crítica es el mismo para cualquier distribución continua F_0 y está tabulado.

Demostración. Para poder probar este lema, necesitamos una proposición:

Proposición 2.4. *Si una $v.a.$ tiene distribución continua F_0, entonces la variable aleatoria $F_0(X) \sim \mathbb{U}(0, 1)$*

[5] El estimador converge al verdadero valor. En particular, acabamos de ver que es consistente *débil* porque converge en *probabilidad*, pero veremos que en realidad tenemos convergencia uniforme.

Demostración.

$$P\{F_0(x) \leq u\} \stackrel{?}{=} u$$

Hemos tomado distribuciones continuas para evitar saltos. Entonces:

F_0 continua $\implies \exists x \nearrow F_0(x) = u, P\{F_0(X) \leq F_0(x)\}$

Donde:

$$\{F_0(X) \leq F_0(x)\} = \underbrace{\{F_0(X) \leq F_0(x), X \leq x\}}_{\{X \leq x\}} \cup \{F_0(X) \leq F_0(x), X \geq x\}$$

Vamos a estudiar cada uno de los términos:

$$\{F_0(X) \leq F_0(x), X \leq x\} = \{X \leq x\}$$

Se debe a que $X \leq x \implies F_0(X) \leq F_0(x)$

Por otro lado, el otro término:

$$\{F_0(X) \leq F_0(x), X \geq x\} = P\{F_0(X) \leq F_0(x), X > x\} = 0$$

Ya que F_0 es necesariamente constante entre x y X.

Entonces:

$$\{F_0(X) \leq F_0(x)\} = P\{X \leq x\} = F(x) = u$$

□

Observación: Existe un recíproco que permite calcular otras distribuciones a partir de uniformes invirtiendo la función de distribución.

Ahora que ya tenemos esa propiedad demostrada podemos seguir con la demostración del lema.

Vamos a demostrar que $D_n = \text{máx}\{\sup_{x\in\mathbb{R}}(F_n(x) - F_0(x)), \sup_{x\in\mathbb{R}}(F_0(x) - F_n(x))\}$

Lo hemos separado en 2 para quitarnos el valor absoluto. Aquí solo estudiaremos uno de los términos, y el otro se deja como ejercicio para el lector.

Por definición de la distribución empírica, tenemos: $x \in [X_{(i)}, x_{(i+1)}), F_n(x) = \frac{i}{n}$

Entonces:

$$\sup_{x\in\mathbb{R}}(F_n(x) - F_0(x)) = \max_{i=0,\dots,n} \sup_{x\in[X_{(i)},X_{(i+1)})} (F_n(x) - F_0(x))$$

Ahora podemos sustituir $F_n(x)$, obteniendo

$$= \max_{i=0,\dots,n} \left(\frac{i}{n} - F_0(x_{(i)})\right) = \text{máx}\left\{0, \max_{i=0,\dots,n}\left(\frac{i}{n} - F_0(x_{(i)})\right)\right\}$$

Por el otro lado obtendríamos:

$$\sup_{x\in\mathbb{R}}(F_0(x)-F_n(x)) = \ldots = \text{máx}\left\{0, \max_{i=0,\ldots,n}\left\{0, \left(F_0(X_{(i)}) - \frac{i-1}{n}\right)\right\}\right\}$$

Conclusión:

$$D_n = ||F_n - F_0||_\infty = \text{máx}\left\{0, \max_{i=0,\ldots,n}\left(\frac{i}{n} - F_0(x_{(i)})\right), \max_{i=0,\ldots,n}\left\{0, \left(F_0(x_{(i)}) - \frac{i-1}{n}\right)\right\}\right\}$$

Es decir, D_n depende de F_0 a través $F_0(X_{(i)})$ y, aplicando la proposición 2.4 en $X_1, \ldots, X_n \overset{iid}{\sim} F_0 \to X_{(1)} \leq \ldots \leq X_{(n)}$

Entonces:

$$F_0(X_1), \ldots, F_0(X_n) \overset{iid}{\sim} \mathbb{U}(0,1) \to F_0(X_{(1)}) \leq \ldots \leq F_0(X_{(n)})$$

Son los estadísticos de de orden de una muestra de tamaño n de v.a.i.i.d $\mathbb{U}(0,1)$ para cualquier F_0 continua.

□

Ejemplo: *Contrastar a nivel α=0.01 si la muestra* $16, 8, 10, 12, 6$ *procede de una exponencial de media* $\lambda = 11.5$

Recordamos que la media de una exponencial: $\mu = \frac{1}{\lambda}$.

En este caso, $F_0(x) = 1 - e^{-\frac{x}{11.5}}$.

Vamos a construir una tabla para resolverlo, llamando $D_n^+ = \{\frac{i}{n} - F_0(X_{(i)})\}$ *y* $D_n^- = \{F_0(X_{(i)}) - \frac{i-1}{n}\}$

$X_{(i)}$	$\frac{i}{n}$	$F_0(X_{(i)})$	D_n^+	D_n^-
6	0.2	0.41	−0.21	0.41
8	0.4	0.5	−0.1	0.3
10	0.6	0.58	0.02	0.18
12	0.8	0.65	0.15	0.05
16	1	0.75	0.25	−0.05

En esta caso, $D_n = \text{máx}\{0, D_n^+, D_n^-\} = 0.41$. *Ahora, como el contraste no depende de la* F_0 *(mientras que esta sea continua), consultamos la tabla y vemos que para* $n = 6$ *y* $\alpha = 0.01$ *obtenemos* $c = 0.669$.

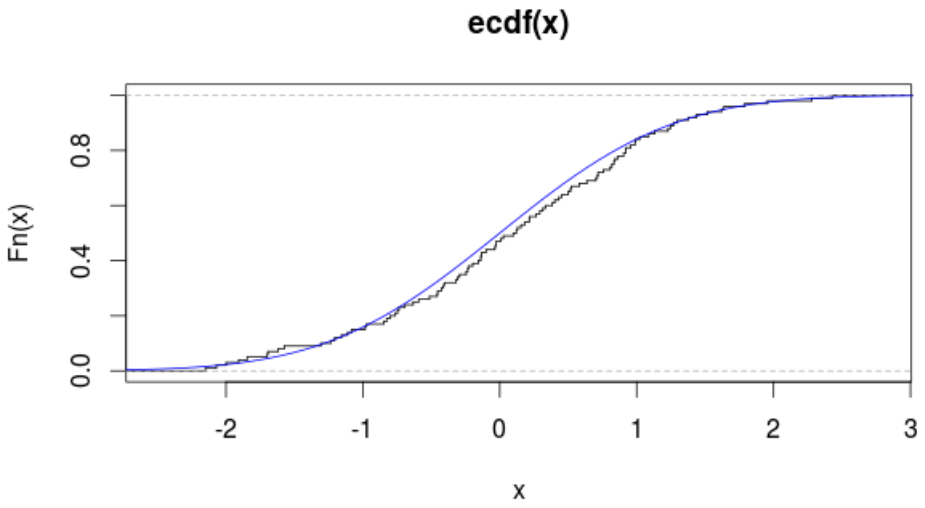
$$c > D_n \iff 0.669 > 0.41$$

Con lo que no podemos rechazar la hipótesis.

2.2.1. Contrastes Kolmogorov-Smirnov con R

Para el ejemplo anterior, el código de R tendríamos algo que aparecerá más adelante:

```
1
2 n = 100
3 x = rnorm(n)
4 plot(ecdf(x),vertical=TRUE, do.points=FALSE)
5 curve(pnorm(x),-4,4add=TRUE,col=blue)
6
7 Dn= ks.test(x,pnorm)$statistic
8 # En esta gráfica, Dn es la mayor distancia vertical entre la negra (empírica
    ) y la azul (real).
9
10
```




```
11
12
13
14 ks.test(x,pnorm) # Cierta. Obtenemos un p-valor grande (del orden de 0.7)
15 ks.test(x,pexp) # Falsa. Obtenemos un p-valor casi 0 (del orden de 10^-16)
```

Iteración del contraste Es bueno repetir el contraste varias veces, ya que la generación de los datos es aleatoria. Para ello, hay un fichero en moodle de R con el proceso. En las transparencias además, está el otro ejemplo.

2.3. Gráficos de probabilidad

Este es otro tipo de contraste. Hemos visto cómo contrastar con una χ^2. También hemos utilizado la distribución empírica.

Ahora vamos a ver cómo contrastar utilizando los cuantiles.

Tenemos:

$$X_1, ..., X_n \overset{iid}{\sim} F \implies F(X_n) \overset{iid}{\sim} \mathcal{U}(0,1) \implies F(X_{(i)}) < F(X_{(i+1)})$$

Podemos calcular:

$$\mathbb{E}\left(F(X_{(i)})\right) = \frac{i}{n+1}$$

Utilizando esta información, podemos dibujar en el eje x de una gráfica los estadísticos de orden $X_{(i)}$ y en el eje y, ponemos: $F^{-1}\left(\frac{i}{n+1}\right)$, ya que, en media $F(X_{(i)}) \sim \frac{i}{n+1}$

Lo que esperaríamos si nuestra hipótesis de F fuera cierta, es que los puntos de esa gráfica estén sobre una recta.

Ejemplo: *Sea* $X_1, ..., X_n \overset{iid}{\sim} \underbrace{N(\mu, \sigma^2)}_{F}$. *Tomamos* $\Phi \sim N(0,1)$

$$X_{(i)} \sim F^{-1}\left(\frac{i}{n+1}\right) = ... = \sigma\Phi^{-1}\left(\frac{i}{n+1}\right) + \mu$$

Si los datos son normales, los puntos $\left(X_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$ *estarán alineados.*

Observación: Este es el tipo de contraste que se utiliza en la realidad para saber si unos datos se distribuyen normalmente o no. Hay una manera de medir esta diferencia y no hacerlo a ojo. Este test se denomina **Shapiro-wilks**

Shapiro-wilks

3. Contrastes de homogeneidad

3.1. χ^2

Tenemos

$$\left.\begin{array}{cc} M_1 \equiv X_{11}, ..., X_{1n_1} & \overset{iid}{\sim} F_1 \\ \vdots & \vdots \\ M_p \equiv X_{p1}, ..., X_{pn_p} & \overset{iid}{\sim} F_p \end{array}\right\} H_0 : F_1 = F_2 = ... = F_p$$

Con $M_1, ..., M_p$, muestras **independientes**.

Proceso para el contraste

1. Dividimos los datos en clases A_i y consideramos las frecuencias observadas $O_{ij} = \{\#\text{ datos } M_j \text{ en } A_i\}$
2. Al considerar número de M_j, tenemos, bajo la hipótesis, que $O_{ij} \equiv B(n_j, p_i)$, con $p_i = P_{H_0}(A_i)$.

Tabla de contingencia

Definición 3.1 **Tabla de contingencia**.

	μ_1	$\cdots$	μ_p
A_1	O_{11}	$\cdots$	O_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_k	O_{k1}	$\cdots$	O_{kp}

3. Las **frecuencias esperadas** bajo $H_0 \to E_{ij} = n_j p_i$.

4. Construimos el estadístico del contraste.

$$\sum_j \sum_i \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow[n\to\infty]{d} \chi^2_{p(k-1)}$$

La independencia de las M_i es la que da la convergencia a la χ^2.

Dado que no conocemos p_i, tenemos que estimarlos a partir de los datos. ¿Cuál es el estimador? Al suponer $F_i = F_j$ podemos utilizar:

$$\hat{p_i} = \frac{\sum_j O_{ij}}{\sum_j n_j} \implies \hat{E}_{ij} = n_j \cdot p_i = \frac{O_{i\cdot}}{n} O_{\cdot j}$$

5. **Conclusión** El estadístico es:

$$T = \sum_i \sum_j \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \xrightarrow[n\to\infty]{d} \chi^2_{(p-1)(k-1)}$$

¿Porqué los grados de libertad? Porque al estimar perdemos un grado de libertad por cada estimación, y como aquí estamos estimando k cosas, perderíamos k grados de libertad. El -1 aparece porque no hace falta estimar k sino $k-1$ ya que $\sum_i \hat{p_i} = 1$. Entonces $p(k-1) - (k-1) = (p-1)(k-1)$

Y la región de rechazo

$$R = \{T > \chi^2_{(k-1)(p-1),\alpha}\}$$

Observación: También podemos construir el estadístico como:

$$T = \sum_i \sum_j \frac{O_{ij}^2}{\hat{E}_{ij}} - n$$

Demostración. Se deja como ejercicio para el lector la afirmación de que el estadístico también se puede construir así. □

Ejemplo: *Tenemos 100 observaciones de 3 países (ESP, ITA, FRA) sobre fumadores y queremos saber si las distribuciones que siguen cada tipo de fumador son iguales.*

	E	*I*	*F*	#
NF	*30*	*15*	*20*	*65*
FO	*50*	*40*	*50*	*140*
FH	*20*	*45*	*30*	*95*
	100	*100*	*100*	*300*

¿Qué esperamos? →

	E	*I*	*F*	#
NF	*21.6*	*21.6*	*21.6*	*65*
FO	*46.6*	*46.6*	*46.6*	*140*
FH	*31.6*	*31.6*	*31.6*	*95*
	100	*100*	*100*	*300*

Ahora podemos calcular el estadístico y la región de rechazo para $\alpha = 0.05$:

$$T = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{O_{ij}^2}{E_{ij}} - n = ... \sim 16.8$$

$$c = \chi^2_{4,0.05} = 9.48$$

Como $T > c \implies$ Rechazamos la hipótesis.

3.2. KS de homogeneidad

Es importante destacar e este tipo de contraste que sólo es válido para **dos muestras** y sólo para **distribuciones continuas**.

Tenemos

$$\left.\begin{array}{l} X_1, ..., X_n \overset{iid}{\sim} F \\ Y_1, ..., Y_m \overset{iid}{\sim} G \end{array}\right\} H_0 : F = G$$

Sean F_n y G_m las funciones de distribución empíricas respectivas.

Calculamos el estadístico de KS para 2 muestras:

$$D_{n,m} = ||F_n - G_m||_\infty = \sup_{a\in\mathbb{R}} |F_n(a) - G_m(a)|$$

Bajo H_0, la distribución de $D_{n,m}$ no depende de $F = G$ y está tabulada (igual que en los contrastes de bondad de ajuste).

Entonces:

$$R = \{D_{n,m} > C_\alpha\}$$

Definición 3.2 **Diferencias estandarizadas al cuadrado**.

Diferencias estandarizadas al cuadrado

$$\frac{(O_{ij} - \hat{E_{ij}})}{\sqrt{\hat{E_{ij}}}}$$

4. Contrastes de independencia

4.1. χ^2

Sean $(X_1, Y_1), ..., (X_n, Y_n) \overset{iid}{\sim} F$ con $H_0 : X, Y$ son independientes.

Los datos se suelen dar en tablas por clases, es decir:

	B_1	$\cdots$	B_p
A_1	O_{11}	$\cdots$	O_{1p}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_k	O_{k1}	$\cdots$	O_{kp}

Una diferencia con el anterior es que que los totales no están fijados, ya que son variables.

Vamos a ver la estimación de las frecuencias esperadas $\hat{E_{ij}}$ y E_{ij}

$$E_{ij} = n \cdot p_{ij} = n \cdot P(x \in A_i, Y \in B_j) \overset{H_0}{=} n \cdot P(x \in A_i)(Y \in B_j)$$
$$\hat{E}_{ij} = n\frac{O_{i\cdot}}{n}\frac{O_{\cdot j}}{n} = \frac{O_{i\cdot} \cdot O_{\cdot j}}{n} \text{ igual que en el anterior}$$

Ahora, calculamos T y R exactamente igual que en el contraste de homogeneidad χ^2 3.1, es decir:

$$T = \sum_i \sum_j \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \xrightarrow[n\to\infty]{d} \chi^2_{(p-1)(k-1)}$$

Recordamos que también podemos construir el estadístico como:

$$T = \sum_i \sum_j \frac{O_{ij}^2}{\hat{E}_{ij}} - n$$

Y la región de rechazo

$$R = \{T > \chi^2_{(k-1)(p-1),\alpha}\}$$

La diferencia entre este contraste y el de homogeneidad es ...

Capítulo II

Regresión

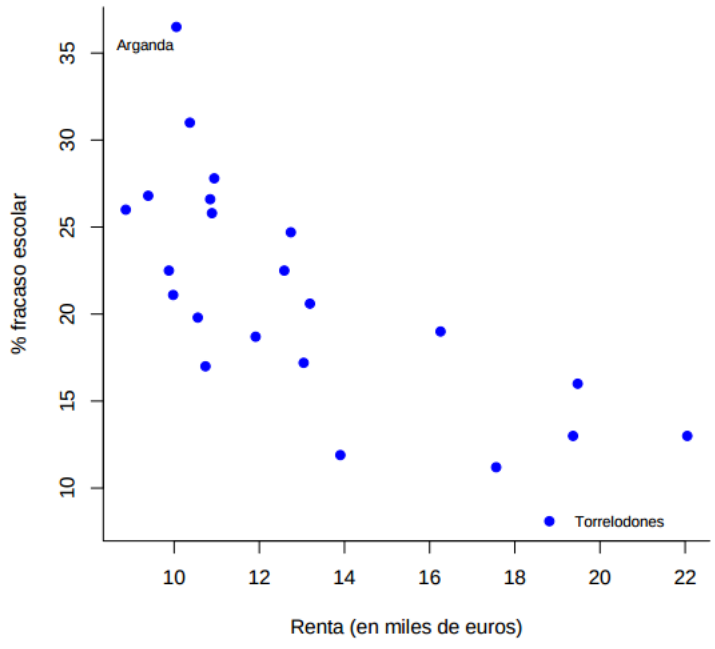
El objetivo de la regresión es predecir una/s variable/s en función de la/s otra/s.

1. Regresión lineal

Observamos dos variables, X e Y , el objetivo es analizar la relación existente entre ambas, de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

- La variable Y se llama variable respuesta.
- La variable X se llama variable regresora o explicativa.

Por ejemplo:



Queremos predecir el fracaso escolar en función de la renta. La variable respuesta es el fracaso escolar, mientras que la variable regresora es la renta.

1.1. Regresión lineal simple

Frecuentemente existe una relación lineal entre las variables. En el caso del fracaso escolar,queremos construir una recta $Y_i = \beta_1 X_i + \beta_0 \; i = 1, ..., n$ que minimice el error.

El problema es estimar los parámetros β_0, β_1. Una manera de hacer esto es:

1.1.1. Recta de mínimos cuadrados

Recta de mínimos cuadrados

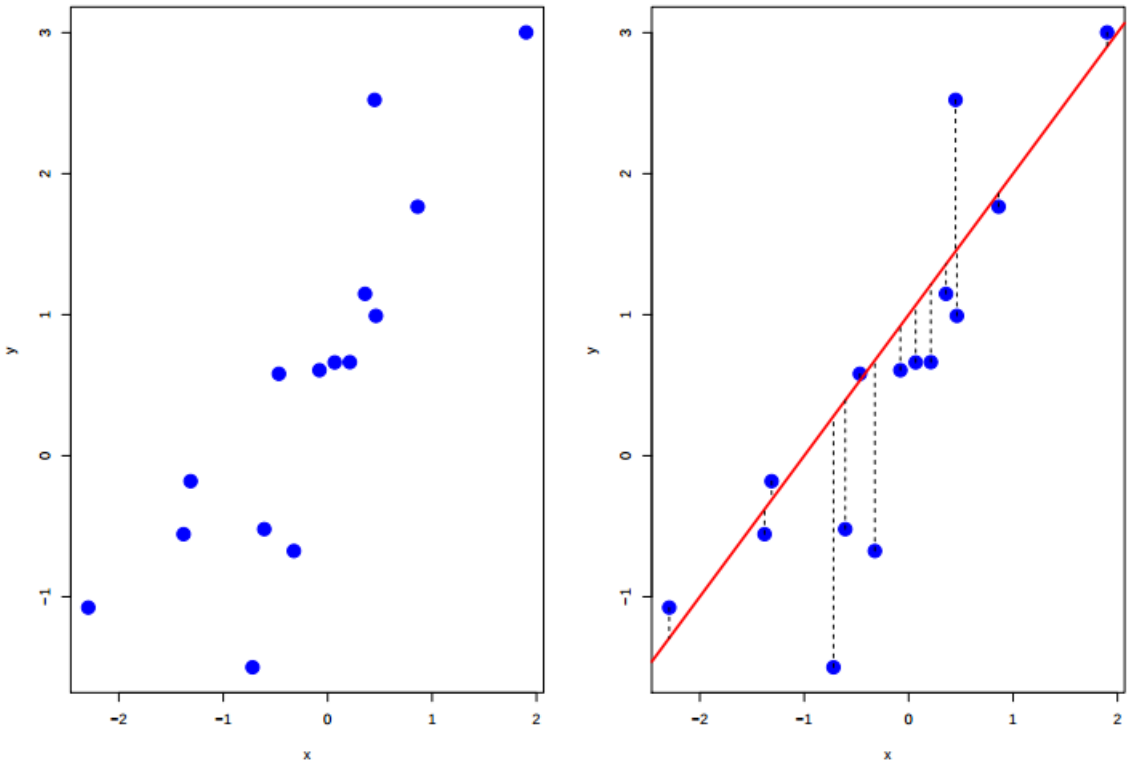
Definición 1.1 **Recta de mínimos cuadrados**. *Estimando β_i por $\hat{\beta}_i$ obtenemos:*

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

La recta viene dada por los valores $\hat{\beta}_0, \hat{\beta}_1$ para los que se minimiza el error cuadrático, es decir:

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2$$

Ejemplo:



Cómo calcular la pendiente de la recta de mínimos cuadrados.

Vamos a ver unas pocas maneras de calcular la recta de mínimos cuadrados.

- El sistema habitual:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Donde

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$$

Y, consecuentemente, como el avispado lector podrá imaginar

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

Es interesante darse cuenta que, siendo S_x la cuasivarianza, tenemos $S_{xx} = (n-1)S_x$

$$\beta_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Entonces:

$$\text{recta} \equiv y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

- Mínimos cuadrados como promedio de pendientes:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{S_{xx}} \left(\frac{(Y_i - \bar{Y})}{x_i - \bar{x}} \right) = \sum_{i=1}^n \omega_i \left(\frac{(Y_i - \bar{Y})}{x_i - \bar{x}} \right)$$

Vemos que hemos ponderado la pendiente de cada recta que une cada punto con la media. Este peso es mayor cuanto mayor es la distancia **horizontal**.

- Mínimos cuadrados como promedio de respuestas:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} \overset{(1)}{=} \sum_{i=1}^n \frac{x_i - \overline{x}}{S_{xx}} Y_i = \sum \alpha_i Y_i$$

$(1) \Longleftarrow$ hemos utilizado una propiedad básica, importantísima y, a simple vista, poco (o nada) intuitiva:

Proposición 1.1. *Sea $\{x_i\}, \{y_i\}$ datos de variables aleatorias.*

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i - \overline{x}) y_i = \sum (y_i - \overline{y}) x_i$$

***Importante:** sólo quitamos la media de una de las 2. No podemos hacer $\sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i$, porque esto ya no es verdad.*

Demostración.

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}(x_i-\overline{x})y_i-\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})\overline{y}}_{0}$$

Vamos a ver por qué ese término es 0.

$$\sum_{i=1}^{n}(x_i-\overline{x})\overline{y}\overset{(1)}{=}\left(\left(\sum_{i=1}^{n}x_i\right)-n\overline{x}\right)\frac{\sum y_i}{n}$$

$(1)\rightarrow$ Estamos restando n veces el término $\overline{x}$ que no tiene índice del sumatorio, con lo que podemos sacarlo fuera.

Aplicando la propiedad distributiva con el factor $\frac{1}{n}$, obtenemos:

$$\left(\frac{\left(\sum x_i\right)}{n}-\frac{n\overline{x}}{n}\right)\sum_{i=1}^{n}y_i=(\overline{x}-\overline{x})\sum y_i=0$$

Observación: Pero... ¿y porqué $\sum(x_i-\overline{x})y_i\neq 0$? ¿Cuál es el fallo de lo siguiente?

$$\sum(x_i-\overline{x})y_i=\frac{\sum(x_i-\overline{x})y_i}{n}\cdot n$$

¿Y aplicamos el mismo razonamiento que antes?

La respuesta es que, en este caso el factor $(x_i-\overline{x})$ está multiplicado por y_i **dentro** del sumatorio, es decir:

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}\left[(x_i-\overline{x})y_i\right]-\sum_{i=1}^{n}\left[(x_i-\overline{x})\overline{y}\right]$$

Y podemos sacar $\overline{y}$ del sumatorio, porque está multiplicando y no tiene índice del sumatorio.

$$\sum_{i=1}^{n}\left[(x_i-\overline{x})y_i\right]-\sum_{i=1}^{n}\left[(x_i-\overline{x})\right]\overline{y}$$

□

Proposición 1.2. *Propiedades de estos α_i*

a. $\sum\alpha_i=0$

Demostración.

$$\sum\alpha_i=\sum\frac{(x_i-\overline{x})}{S_{xx}}=0\iff\sum(x_i-\overline{x})=0$$

$$\sum(x_i-\overline{x})=\left(\sum x_i\right)-n\overline{x}=\left(\sum x_i\right)-n\frac{\sum x_i}{n}=0$$

□

b. $\sum \alpha_i x_i = 1$

Demostración.

$$\sum \alpha_i x_i = \sum \frac{(x_i - \overline{x})x_i}{S_{xx}} = \frac{1}{S_{xx}} \sum (x_i - \overline{x})(x_i - \overline{x}) = \frac{S_{xx}}{S_{xx}} = 1$$

□

c. $\sum \alpha_i^2 = \frac{1}{S_{xx}}$

Demostración.

$$\sum \alpha_i^2 = \sum \frac{(x_i - \overline{x})(x_i - \overline{x})}{S_{xx}^2} = \sum \frac{(x_i - \overline{x})x_i}{S_{xx}^2} = \sum \frac{\alpha_i x_i}{S_{xx}} = \frac{1}{S_{xx}} \sum \alpha_i x_i$$

Utilizando el anterior, tenemos $\sum \alpha_i^2 = \frac{1}{S_{xx}}$ □

Residuo

Definición 1.2 **Residuo**. *En una recta de mínimos cuadrados: Sea* $y_i = \beta_1 x_i - \beta_0$ *y sea* $\hat{y}_i = \hat{\beta}_1 x_i - \hat{\beta}_0$*, llamamos residuo a*

$$e_i = y_i - \hat{y}_i$$

Los residuos cumplen:

$$\sum_{i=1}^{n} e_i = 0$$

Esto es intuitivo, ya que los errores se compensan y además es una buena propiedad.

Demostración. En la recta de mínimos cuadrados calculamos $\hat{\beta}_0, \hat{\beta}_1$ con la finalidad de minimizar $\sum (y_i - \beta_0 - \beta_1 x_i)^2$. Si recordamos, derivando respecto de β_0 obteníamos:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^{n} y_i - \beta_0 - \beta_1 x_i$$

Y definimos $\hat{\beta}_0, \hat{\beta}_1$ como los parámetros que igualaban dicha derivada a 0 (minimizaban la suma de diferencias al cuadrado), es decir $\hat{\beta}_0, \hat{\beta}_1$ cumplen:

$$\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$$

Pero precisamente $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ y por tanto tenemos que:

$$\sum_{i=1}^{n} y_i - \hat{y}_i = \sum_{i=1}^{n} e_i = 0$$

□

Proposición 1.3. *Sean* $\{e_i\}$ *una variable aleatoria que cumple* [1]*:*

$$\sum e_i = 0$$

[1] Se ha utilizado la e porque es útil en cuanto a los residuos de la recta de mínimos cuadrados

Entonces:

$$\sum e_i x_i = 0 \implies \text{cov}\,(e, x) = 0$$

Demostración.

$$\text{cov}\,(e, x) = \mathbb{E}\,(e)\,\mathbb{E}\,(x) - \mathbb{E}\,(e \cdot x)$$

Vamos a ver que los 2 sumandos son 0.

$\mathbb{E}\,(e)\,\mathbb{E}\,(x) = 0 \impliedby \mathbb{E}\,(e) \stackrel{?}{=} \overline{e} = 0$

Por otro lado:

$$\sum (e_i - \mu) x_i = \sum (e_i - \mu)(x_i - \mathbf{x})$$

$$\mathbb{E}\,(e \cdot x) = \sum e_i x_i = \sum e_i x_i - \mathbf{x} \sum e_i = \sum e_i (x_i - \mathbf{x})$$

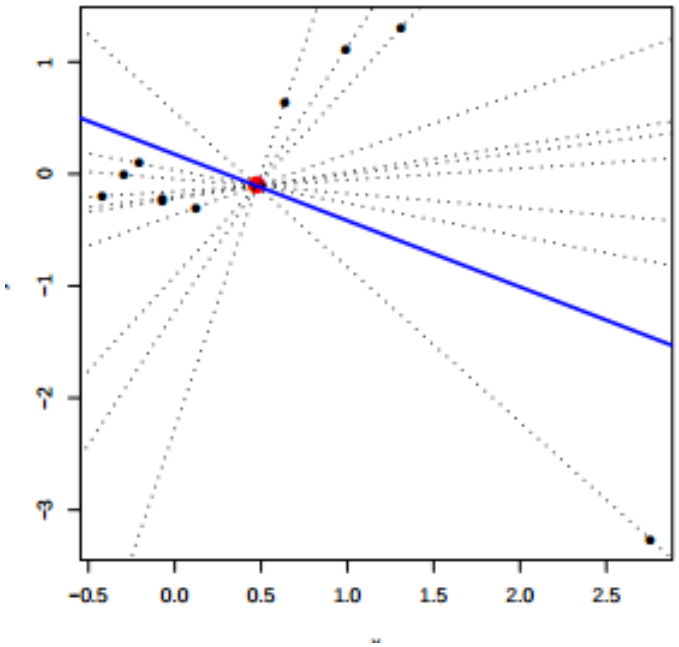
□

Esto tiene la siguiente explicación "intuitiva": La recta de mínimos cuadrados contiene toda la información lineal que X puede dar sobre Y (debido a que la covarianza entre los residuos y X es 0).

1.1.2. Fallos de la recta de mínimos cuadrados

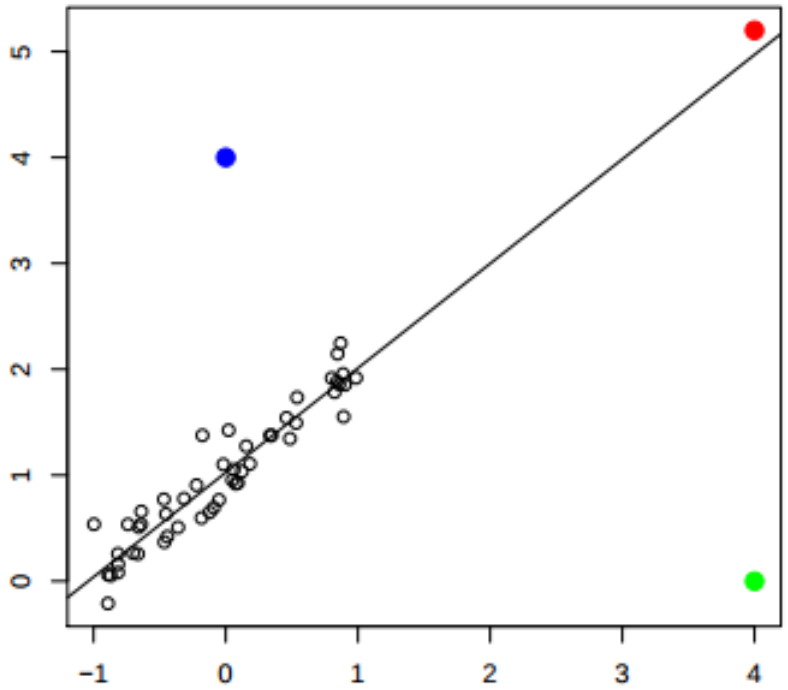
Vamos a ver un par de ejemplos ilustrativos:

Ejemplo: Sobre los datos atípicos *Esta es una recta de mínimos cuadrados calculada para una nube de puntos a la que se ha añadido un punto atípico. Se ve una cierta tendencia de que la pendiente debería ser positiva, pero el dato atípico provoca un cambio brusco.*

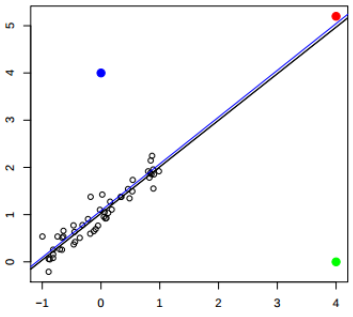


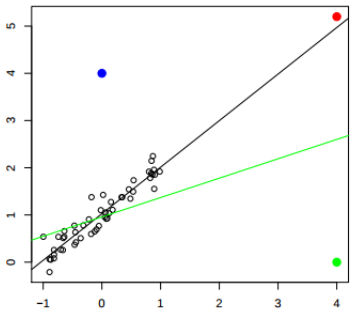
Ejemplo: Sobre la distancia horizontal *¿Y da igual lo atípico que sea un dato? La respuesta es que no. Si el dato es muy atípico en la variable respuesta (Y), pero es muy típico en la variable regresora, la recta no se desvía tanto. Vamos a verlo y después explicamos la razón.*

Esta es la recta, en la que hemos ignorado los 3 datos que parecen "atípicos".



Ahora calculamos las rectas teniendo en cuenta sólo uno de los puntos.





Vemos que la recta azul no se desvía apenas de la original, mientras que la recta verde si se desvía un montón. ¿Esto a qué se debe? A que importa más la distancia horizontal de la media que la distancia vertical. Si vamos a la expresión de la recta de mínimos cuadrados como promedio de las pendientes 1.1.1 vemos que hay un término $\frac{(x_i-\overline{x})}{S_{xx}}$ que hemos tomado como pesos para ponderar y en este caso, la distancia horizontal $(x_i - \overline{x})$ está multiplicando en el numerador.

1.1.3. Introduciendo "aleatoreidad" para poder hacer IC

Sea $\{\varepsilon_i\}$ siendo $\varepsilon_i \sim N(0, \sigma^2)$. Lo habitual es no saber cómo han sido generados los datos y es probable que no vayamos a conocer con exactitud absoluta la recta de mínimos cuadrados. Es por ello que suponemos el siguiente modelo para la variable respuesta:

$$Y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$$

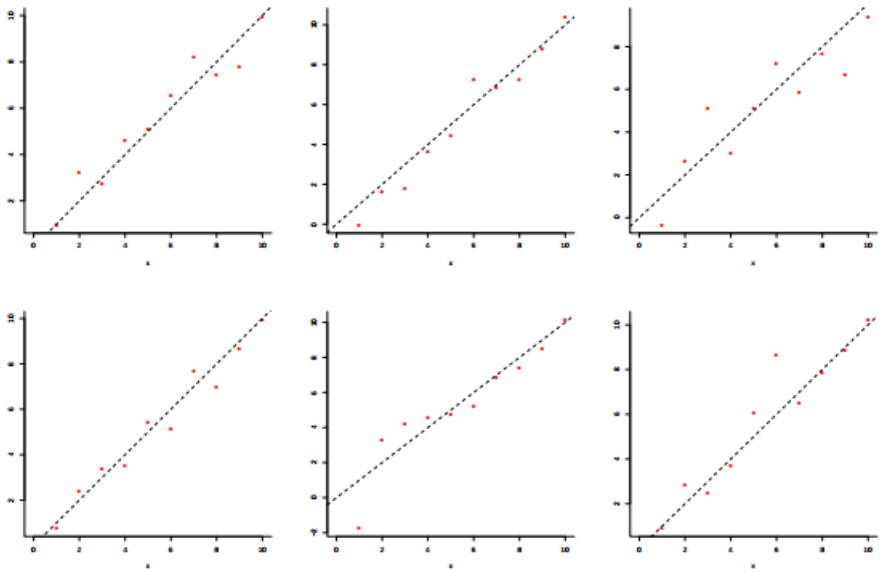
Tenemos que $\bar{y}_i \sim N$, ya que es una combinación lineal de variables normales **independientes** (como vimos en el Tema 1).

Ejemplo: *Sea $\sigma = 1, \beta_0 = 0$ y $\beta_1 = 1$.*

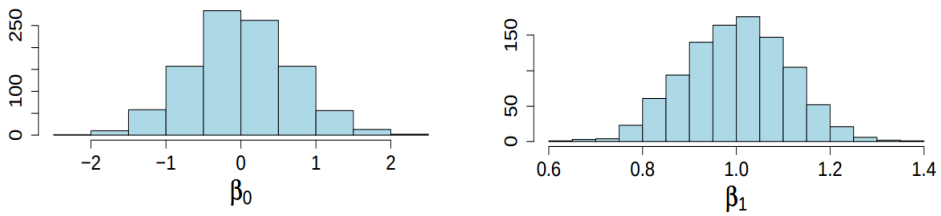
Entonces el modelo es:

$$Y_i = x_i + \varepsilon_i$$

Fijamos $n = 10$ y generamos las respuestas para $x_i = i$. Además, repetimos el experimento 6 veces y calculamos las rectas de mínimos cuadrados, obteniendo:



Vemos que obviamente las rectas no son las mismas. Esto se debe al ε_i introducido. ¿Cuáles son los valores que toman β_1 y β_0? Habiendo repetido el experimento 1000 veces, obtenemos los siguientes histogramas:



Vemos que no siempre es el mismo valor. Sabemos (por cómo hemos construido los datos) que $\beta_0 = 0$ y $\beta_1 = 1$, pero nuestra manera de calcularos (debido a ε_i) no siempre nos da el valor concreto.

El ejemplo anterior nos muestra que en realidad, estamos estimando β_i, aunque no nos guste y ahora tenemos que plantearnos ¿Cómo de buenos son nuestros estimadores? Tal vez son una mierda, o tal vez son insesgados.

Para ello, vemos que al haber añadido un error $\varepsilon_i \sim N(0, \sigma^2)$, tenemos:

$$Y_i = \beta_0 + \beta_1 x + \varepsilon_i \implies Y_i \equiv N(\beta_0 + \beta_1 X_i, \sigma^2)$$

1.1.4. Estimando β_1

Proposición 1.4. *Nuestro estimador "pendiente de la recta de mínimos cuadrados:" $\hat{\beta}_1$ cumple*

$$\hat{\beta}_1 \equiv N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Demostración.

- $\mathbb{E}\left(\hat{\beta}_1\right) = \sum_{i=1}^n \alpha_i \mathbb{E}(Y_i) = \sum_{i=1}^n \alpha_i(\beta_0 + \beta_1 x_i) = \beta_0 \underbrace{\sum \alpha_i}_{=0} + \beta_1 \underbrace{\sum \alpha_i x_i}_{=1} = \beta_1$
- $\mathbb{V}\left(\hat{\beta}_1\right) = \mathbb{E}(\hat{\beta}_1^2) - \mathbb{E}^2(\hat{\beta}_1) = \mathbb{V}\left(\sum_{i=1}^n \alpha_i y_i\right) \underbrace{=}_{y_i \text{ independientes}} \sum \alpha_i^2 \sigma_i^2 \underbrace{=}_{\text{homocedasticidad}} \frac{\sigma^2}{S_{xx}}$

□

1.1.5. Estimando β_0

Proposición 1.5. *Nuestro estimador "término independiente de la recta de mínimos cuadrados:" $\hat{\beta}_0$ cumple*

$$\hat{\beta}_0 = N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$$

Demostración.

- $\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0$
- $\mathbb{V}\left(\hat{\beta}_0\right) = \mathbb{V}\left(\overline{Y}\right) + \mathbb{V}\left(\hat{\beta}_1 \overline{x}\right) - 2 \cdot \operatorname{cov}\left(\overline{Y}, \hat{\beta}_1 \overline{x}\right) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} - 2\overline{x} \cdot \operatorname{cov}\left(\overline{Y}, \hat{\beta}_1\right)$

  Calculamos: $\operatorname{cov}\left(\overline{Y}, \hat{\beta}_1\right)$ utilizando cosas del tema 1

  $$\operatorname{cov}\left(\overline{Y}, \hat{\beta}_1\right) = \operatorname{cov}\left(\frac{1_n' Y}{n}, \alpha' Y\right) = \frac{1}{n} 1_n' \sigma^2 I \alpha = 0$$

  debido a que $\alpha = 0$.

Ademas de ser incorrelados, son **independientes**. ¿Por qué? Porque conjuntamente son normales, es decir

$$\begin{pmatrix} \overline{Y} \\ \hat{\beta}_1 \end{pmatrix} \equiv A\overline{Y} \equiv N_2$$

□

Conclusiones:

$$\overline{Y} \text{ es independiente de } \hat{\beta}_1$$
$$\hat{\beta}_1 \equiv \left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$
$$\hat{\beta}_0 \equiv \left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right)$$

¿Son estas las variables $\hat{\beta}_1$ y $\hat{\beta}_0$ normales una normal conjunta? Sí, **sí son una normal conjunta**. Una manera que tenemos de saber si es una normal conjunta es si son independientes, y en este caso no lo son. Intuitivamente es fácil de ver. En una recta, si aumentamos la pendiente (y estamos en el primer cuadrante) entonces el término independiente disminuye.

Esta dependencia tiene que aparecer. Vamos a estudiar la covarianza entre los estimadores:

$$\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \text{cov}\left(\overline{Y} - \hat{\beta}_1\overline{x}, \hat{\beta}_1\right) = \text{cov}\left(\overline{Y}, \hat{\beta}_1\right) - \overline{x}\text{cov}\left(\hat{\beta}_1, \hat{\beta}_1\right) = -\overline{x}\frac{\sigma^2}{S_{xx}}$$

Pero sabemos que sí son una normal bidimensional porque toda combinación lineal de nuestros parámetros de la recta es una variable aleatoria (la variable regresora $\hat{Y}$) normal. Es decir:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = AY$$

1.1.6. IC y Contrastes para β_1

Recordamos que

$$\hat{\beta}_1 \equiv N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Podemos normalizar y buscar una cantidad pivotal (como hacíamos en estadística I)

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \equiv N(0,1)$$

Pero aquí nos encontramos con que necesitamos σ, la varianza de los errores. Esta varianza a menudo no es conocida (porque no sabemos con exactitud cuál es la recta verdadera) y tenemos que estimarla.

Para estimarla, parece razonable usar

$$\hat{\sigma} = S_R = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Explicación: Recordamos que para que estimar la varianza, utilizamos (por el lema de Fisher) $n-1$ de denominador para que el estimador sea insesgado. Esto sale de que en la demostración, hay una matriz de rango $n-1$ ya que existe una restricción.

Siguiendo este razonamiento, en este caso tenemos 2 restricciones[2], por lo que si lo demostráramos rigurosamente, aparecería una matriz de rango $n-2$ y por eso es el denominador. De esta manera, conseguimos un estimador insesgado.

Varianza residual

Además, S_R se denomina **Varianza residual**

Proposición 1.6. *Una pequeña generalización del lema de Fisher:*

$$\frac{(n-2)S_R^2}{\sigma^2} \equiv \chi^2_{n-2}$$

Además, es independiente de $\hat{\beta}_1$

Demostración. Esta proposición es un caso particular de un teorema que veremos más adelante. □

Llamamos

$$P_R = \frac{\hat{\beta}_1 - \beta_1}{\frac{S_R}{\sqrt{S_{xx}}}}$$

$$P_\sigma = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}$$

Sabemos que $P_\sigma \sim N(0,1)$. Pero, al estimar ¿Se mantiene$P_R \sim N(0,1)$?

Al estimar σ, P_R no tiene porqué ser exactamente $N(0,1)$. Si $n \to \infty$, entonces $S_R = \sigma$ y entonces $P_\sigma = P_R = N(0,1)$, pero para otros valores de $n \neq \infty$...

Por ello, nos vemos en la necesidad de manipular P_R algebraicamente a ver si conocemos qué distribución tiene (que debería ser algo cercano a una normal, ya que estamos estimando σ con un estimador insesgado. Tal vez las colas de la distribución son menos pesadas y podríamos esperar que fuera una t)

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{S_R}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}\frac{S_R}{\sigma}} = \left(\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}\right)\frac{1}{\frac{S_R}{\sigma}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}}}{\frac{S_R}{\sigma}}$$

En el numerador tenemos una $N(0,1)$ y en el denominador la raíz de una χ^2 dividida por sus grados de libertad (por la proposición anterior). Esto es por definición una t (T-Student) con los mismos grados de libertad que la χ^2, es decir $n-2$. (Wikipedia)

Proposición 1.7. *Podemos calcular el intervalo de confianza para la pendiente de la recta, utilizando la fórmula de intervalo de confianza*

$$IC_{1-\alpha}(\beta_1) \equiv \left[\hat{\beta}_1 \mp t_{n-2,\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}}\right] = \left[\hat{\beta}_1 \mp \mathcal{Z} \frac{\sigma}{\sqrt{S_{xx}}}\right]$$

1.1.7. Contrastes de tipo $H_1 : \beta_1 > 0$

En este tipo de contrastes ponemos como hipótesis nula lo contrario a lo que queremos afirmar. Imaginemos que queremos afirmar que $\beta_1 > 0$, en dicho caso esa es la hipótesis alternativa mientras que la hipótesis nula es $H_0 : \beta_1 \leq 0$. De este modo la región de rechazo a nivel α queda definida como:

$$R = \left\{ \frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} > t_{n-2;\alpha} \right\}$$

ya que lo que queremos (hipótesis alternariva $H_1 : \beta_1 > 0$) es que β_1 sea suficientemente positivo.

Observación: Ojo que es t_α no $t_{\frac{\alpha}{2}}$ como habitualmente.

1.1.8. Contraste en R

```
# Ajusta el modelo
regresion = lm(Fracaso~Renta)
summary(regresion)

lm(formula = Fracaso ~ Renta)

Residuals:   Min    1Q       Median     3Q     Max
          −7.8717 −3.7421    0.5878   3.0368   11.5423
——
Coefficients: Estimate Std. Error    t−value    Pr(>|t|)
(Intercept)    38.4944        3.6445  10.562     8.37e−10 ***
Renta          −1.3467        0.2659  −5.065     5.14e−05 ***
——
Signif. codes: [...]
Residual standard error: 4.757 on 21 degrees of freedom
Multiple R−Squared: 0.5499,
Adjusted R−squared: 0.528
```

Aquí, la fila de *intercept* es el término independiente y renta es la pendiente. Además, los p-valores son para el contraste $\hat{\beta}_i = 0$, dentro de la hipótesis $\beta_i \geq 0$. [3].

En este caso, el p-valor para $H_0 : \hat{\beta}_1 =$ es $5.14e-5$, con lo que no podemos rechazar la hipótesis.

1.1.9. Predicciones

Sea $(x_1, y_1), ..., (x_n, y_n) \rightarrow y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Dado una nueva observación x_0, tenemos 2 problemas para predecir:

[3] Si queremos contrastar si es positivo, nos vamos al caso límite que lo separa y contrastamos eso

- **Inferencia sobre** $m_0 \equiv \mathbb{E}\left(y_0|x_0\right) = \beta_0 + \beta_1 x_0$

  En este caso,

  $$\hat{m}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

  ¿Cómo es este estimador?

  $$\mathbb{E}\left(\hat{m}_0\right) = \beta_0 + \beta_1 x_0 = m_0$$
  $$\mathbb{V}\left(\hat{m}_0\right) = ... = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

  Intuitivamente, lo que significa el segundo sumando de la varianza es que "cuanto más cerca esté x_0 de la media, mejor será la estimación".

  Conclusión:

  $$\hat{m}_0 \sim N\left(m_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

  Intervalo de confianza para m_0 utilizando la fórmula de intervalos de confianza:

  $$IC_{1-\alpha}(m_0) \equiv \left[\hat{m}_0 \pm t_{n-2,\frac{\alpha}{2}} S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right]$$

- **Predecir** Y_0 usamos de nuevo:

  $$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow Y_0 - Y \equiv N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\right)$$

  Donde la varianza ha sido calculada:

  $$\mathbb{V}\left(Y_0 - \hat{Y}_0\right) = \underbrace{\mathbb{V}\left(Y_0\right)}_{\sigma^2} - \underbrace{\mathbb{V}\left(\hat{Y}_0\right)}_{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)} + \underbrace{2\text{cov}\left(Y_0, \hat{Y}_0\right)}_{=0 \text{ (indep.)}} = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

  Este es un problema más complicado, ya que tenemos que tener en cuenta el término de error ε_i y es por esto que aparece el 1 en la varianza. Tenemos que tener en cuenta la incertidumbre.

  Estandarizando y cambiando σ por S, tenemos:

  $$\frac{Y_0 - \hat{Y}_0}{S_R\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \equiv t_{n-2}$$

  Ya que tenemos una normal estandarizada dividida por su .... que por definición, es una t de student.

Intervalo de predicción

Ahora, vamos a construir el **intervalo de predicción** (cambia ligeramente la inter-

pretación)

$$1-\alpha = P\left\{-t_{n-2;\frac{\alpha}{2}} < \frac{Y_0 - \hat{Y}_0}{ET(Y_0 - \hat{Y}_0)} < t_{n-2;\frac{\alpha}{2}}\right\} = P\left\{Y_0 \in \left[\hat{Y}_0 \pm t_{n-2;\frac{\alpha}{2}} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right]\right\}$$

Ahora vamos a hacer unos ejemplos numéricos.

Ejemplo: *Seguimos con el ejemplo de la renta.*

***Ejercicio 1.1**:*

	media	*desviación típica*
% fracaso	*20.73*	*6.927*
renta	*13.19*	*3.814*

La renta está expresada en miles de euros.

***a)** IC para β_1 de nivel $95\,\%$.*

***b)** IC para % de fracaso medio si la renta es de 14.000 euros.*

Es parte del enunciado la salida de "R" incluida en: 1.1.8

APARTADO A)

$$IC_{1-\alpha}(\beta_1) = \left[-1.3467 \mp t_{21;0.025} \cdot (0.2659)\right]$$

Donde el -1.3467 es el estimador $\overline{\beta_1}$ que obtenemos de la salida de R. Lo mismo el 0.2659, que es el error típico.

APARTADO B)

$$\overline{Y_0} = 38.49 - (1.3467) \cdot \underbrace{14}_{x_0} = 19.64$$

Siendo este el estimador, vamos a construir el intervalo de confianza. [4]

$$IC_{1-\alpha}(m_0) = \left[19.64 \mp (2.06)(4.757)\sqrt{\frac{1}{23} + \frac{(14 - 13.19)^2}{S_{xx}}}\right]$$

Donde $S_{xx} = 320.06$ y podemos calcularlo despejando de:

$$E.T.(\overline{\beta_1}) = \sqrt{\frac{S_R^2}{S_{xx}}} \to \sqrt{S_{xx}} = \frac{4.757}{0.2659} \to S_{xx} = 320.06$$

[4] Podría ser que nos pidieran el intervalo de predicción, pero en ese caso estarían pidiendo el intervalo de ..... para predecir. Además, nos están dando un x_0 que para predicción no lo tenemos

Donde $E.T.(\widehat{\beta_1})$ es el error típico dado por R. En este caso es 0.2659y $S_R^2 = 4.757^2$

También podríamos utililzar $S_x = \frac{S_{xx}}{n-1}$, sabiendo que $S_x^2 = \frac{n}{n-1}\sigma^2$. Sabemos que $S_x = 3.814$ por ser la renta la variable explicativa, es decir, las x de nuestro modelo de regresión.

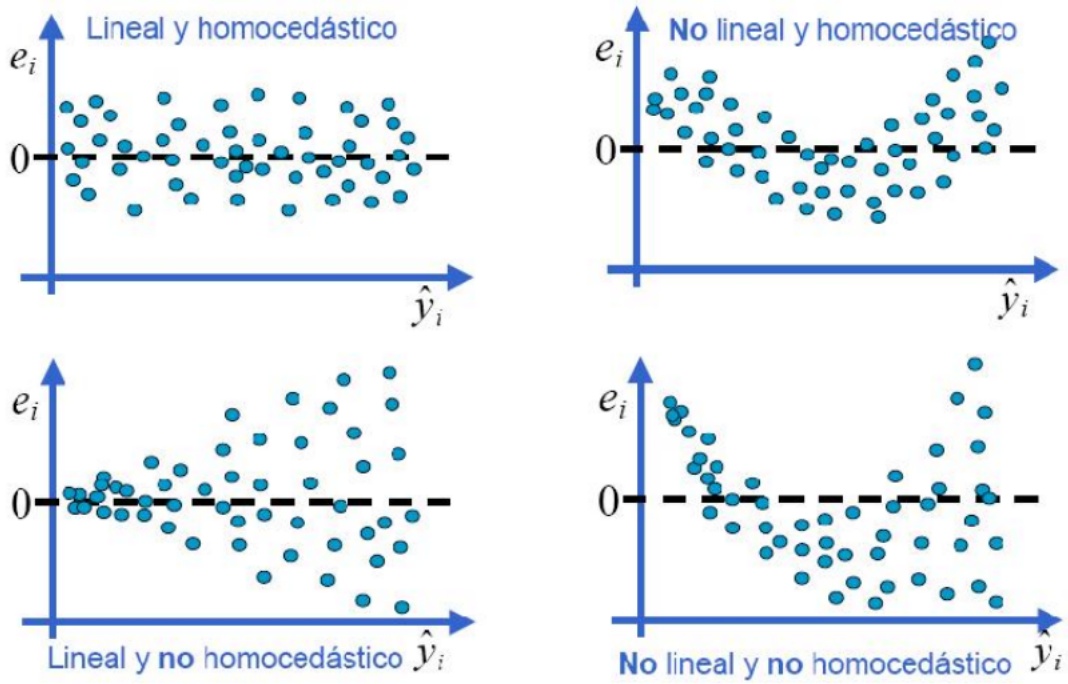
$$\frac{n}{n-1}(3.814)^2 = \frac{S_{xx}}{n-1} \rightarrow S_{xx} = 21 \cdot \left(3.814^2 \cdot \frac{22}{21}\right) = 320.03$$

Diagnóstico del modelo

Observación: Todos estos cálculos y todas estas fórmulas se basan en muchas hipótesis (como que la distribución del error sigue una distribución normal). Pero podría ser que esto no ocurriera y estuviéramos suponiendo un modelo falso. Para ello, en estadística existe el **Diagnóstico del modelo**. Este diagnóstico, consiste en comprobar si las hipótesis del modelo son **aceptables** para los datos disponibles. ¡Ojo! Aceptable... Puede haber muchos modelos aceptables para un mismo conjunto de datos.

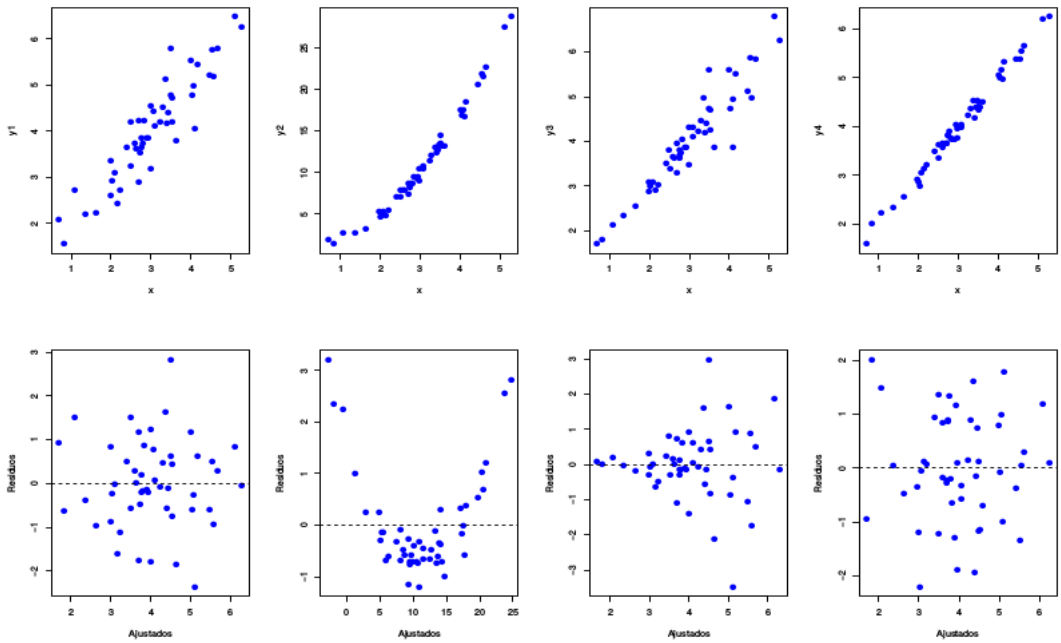
Este diagnóstico se suele basar en el análisis de los residuos del modelo.

Ejemplo: *Vamos a ver a ojo unos cuantos ejemplos. Vamos a utilizar que $\text{corr}(e, \overline{y}) = 0$ bajo el modelo (como calculamos anteriormente)*



De estos 4 gráficos, el bueno es el primero, ya que los demás no cumplen alguno.

Ejemplo: *Vamos a ver otro ejemplo, donde arriba están los datos y abajo los residuos. Mirando sólo la fila de arriba podríamos saber si nuestro modelo para la regresión se cumple o sino.*



Vemos que el primero y el último si tienen este modelo como aceptable, ya que en los residuos no hay ningún patrón (y se cumple que la correlación es 0).

En el segundo, podríamos suponer que es bueno, pero al diagnosticar el modelo mirando los residuos, vemos que no. El diagnóstico del modelo ***magnifica los errores****.*

En el cuarto, vemos más claro que es heterocedástico y que no se cumple el modelo supuesto.

En regresión múltiple veremos que no podemos ver los datos, ya que son demasiadas variables, pero sí podemos estudiar los residuos como acabamos de hacer en los ejemplos anteriores.

1.2. Regresión lineal múltiple

El ejemplo que vamos a estudiar en regresión múltiple es el consumo de gasolina en EEUU intentando predecirlo a partir de unas cuantas variables. Las variables regresoras son:

State	Drivers	FuelC	Income	Miles	MPC	Pop	Tax
AL	3559897	2382507	23471	94440	12737.00	3451586	18.0
AK	472211	235400	30064	13628	7639.16	457728	8.0
AZ	3550367	2428430	25578	55245	9411.55	3907526	18.0

Estos son los datos que obtenemos de *R*:

```
reg <- lm(FuelC ~ Drivers+Income+Miles+MPC+Tax, data=fuel2001)
summary(reg)

Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.844e+05  8.102e+05  -0.598 0.552903
Drivers      6.144e-01  2.229e-02  27.560  < 2e-16 ***
Income       7.526e+00  1.611e+01   0.467 0.642587
Miles        5.813e+00  1.587e+00   3.664 0.000652 ***
MPC          4.643e+01  3.488e+01   1.331 0.189820
Tax         -2.114e+04  1.298e+04  -1.629 0.110298
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394100 on 45 degrees of freedom
Multiple R-squared:  0.9808,    Adjusted R-squared:  0.9787
F-statistic: 459.5 on 5 and 45 DF,  p-value: < 2.2e-16
```

1.2.1. Notación

- n es el número de observaciones, en este caso, el número de estados.
- k es el número de atributos.
- $\varepsilon_i \sim N(0, \sigma^2)$
- $n \geq k + 2$: esta hipótesis es muy necesaria.[5]

Regresión simple es un caso particular de múltiple, tomando $k = 1$.

1.2.2. Modelo

Tenemos una muestra de n observaciones de las variables Y y $X_1, \ldots, X_k$. Para la observación i, tenemos el vector $(Y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$.

El modelo de regresión lineal múltiple supone que:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = 1, ..., n$$

donde las variables de error ε_i verifican:

1. ε_i tiene media cero, para todo i.
2. Var(ε_i) = σ^2, para todo i (homocedasticidad).
3. Son variables independientes.
4. Tienen distribución normal.
5. $n \geq k + 2$ (hay más observaciones que parámetros).
6. Las variables X_i son linealmente independientes entre sí (no hay colinealidad).

[5] En la estadística, habría que rehacer el modelo para cuando $k > n$. ¿Y cuándo $k > n$? ¿Cuándo puede ocurrir esto? Cada vez más hay más información para cada individuo. En estudios genéticos por ejemplo, que hay millones de genes pero no se pueden hacer el estudio con millones de personas... **LA MALDICIÓN DE LA DIMENSIONALIDAD** que decimos en Introducción previa a los Fundamentos Básicos del Aprendizaje Automático.
Una posible solución al problema es un algoritmo que filtre los atributos que son importantes.

Las 4 primeras hipótesis se pueden reexpresar así: las observaciones de la muestra son independientes entre sí con

$$Y_i \equiv N(\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \sigma),\ i = 1, ..., n$$

En forma matricial:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

De forma más compacta, el modelo equivale a:

$$Y = X\beta + \varepsilon,\ \varepsilon \equiv N_n(0, \sigma^2 I_n) \iff Y \equiv N_n(X\beta, \sigma^2 I_n)$$

Matriz de diseño

Definición 1.3 **Matriz de diseño**. *La matriz X se conoce como matriz de diseño*

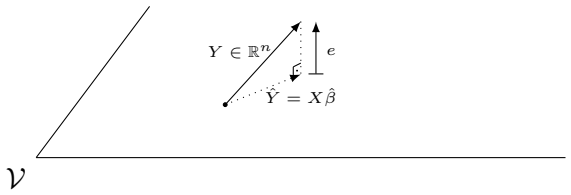
1.2.3. Estimación de los parámetros del modelo

La pregunta que lógicamente se nos viene a la cabeza en este momento es: ¿Cómo estimar β a partir de Y y X?. Para ello nos serviremos de la interpretación geométrica del modelo:

Sea $\mathcal{V} \subset R^n$ el subespacio vectorial generado por las columnas de la matriz de diseño X ($\dim(\mathcal{V}) = k + 1$).

$$\mu \in \mathcal{V} \iff \exists \beta \in R^{k+1} : \mu = X\beta$$

El modelo equivale a suponer $Y \equiv N_n(\mu, \sigma^2 I_n)$, donde $\mu \in \mathcal{V}$.



Con esto, parece razonable estimar μ mediante la proyección ortogonal de Y sobre $\mathcal{V}$ para obtener $\hat{Y} = X\hat{\beta}$. Equivalentemente: $\left\|Y - X\hat{\beta}\right\|^2 \leq \|Y - X\beta\|^2, \forall \beta \in \mathbb{R}^{k+1}$

Estimador mínimos cuadrados

Definición 1.4 **Estimador mínimos cuadrados**.

Al $\hat{\beta}$ que minimiza

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

se le denomina estimador de mínimos cuadrados.

Veamos qué podemos sacar de lo visto hasta ahora para averiguar quién es exactamente $\hat{\beta}$:

Para que $\hat{Y}$ sea la proyección de Y sobre $\mathcal{V}$ es necesario y suficiente que se satisfagan las ecuaciones normales:

Ecuaciones normales

Definición 1.5 **Ecuaciones normales**. *Tomando $e = Y - \hat{Y}$ como el vector residuo, las ecuaciones normales se satisfacen si:*

$$X'(Y - \hat{Y}) = 0 \iff X'e = 0$$

Que se satisfagan estas ecuaciones es equivalente a decir que la resta $Y - \hat{Y}$ da un vector perpendicular a la base $\mathcal{V}$. Sustituyendo $\hat{Y} = X\hat{\beta}$:

$$X'(Y - X\hat{\beta}) = 0 \iff X'Y = X'X\hat{\beta}$$

Ya que las filas de X' (las columnas de X) son independientes, sabemos que $X'X$ tiene rango completo y por tanto es invertible. Y llegamos a la expresión para nuestro estimador de mínimos cuadrados $\hat{\beta}$:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y} \tag{1.1}$$

Observación: En regresión simple se tiene que:

$$X \equiv \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \text{ y que: } X'X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

Con lo visto hasta el momento sabemos que $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$, es decir, que nuestra $\hat{Y}$ está expresada como producto de Y por una matriz que llamaremos:

Hat matrix

Definición 1.6 **Hat matrix**.

$$H = X(X'X)^{-1}X'$$

se conoce como hat matrix, puesto que da Y gorro: $\hat{Y} = HY$.

La hat matrix H tiene las siguientes **propiedades**:

- Es matriz de proyección ortogonal sobre $\mathcal{V}$.
- Es simétrica e idempotente.
- Tiene rango $k+1$ (la dimensión del espacio $\mathcal{V}$ sobre el que proyecta).

Observación: Podemos servirnos de la hat matrix para expresar el vector de residuos:

$$e = Y - \hat{Y} = Y - HY = (I - H)Y$$

Donde $(I-H)$ es una matriz simétrica e idempotente con rango $rg(I-H) = n-(k+1)$, que proyecta sobre el espacio ortogonal $\mathcal{V}^{\perp}$.

Para acabar esta sección enumeramos algunas propiedades de los parámetros:

- $\hat{\beta}$ es el estimador de máxima verosimilitud (EMV) de β:

$$L(\beta, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left\{-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right\}.$$

- El EMV de σ^2 es:

$$\hat{\sigma}^2 = \frac{\|Y - X\beta\|^2}{n} = \frac{\|e\|^2}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

- El vector $\hat{\beta}$ tiene distribución $N_{k+1}(\beta, \sigma^2(X'X)^{-1})$ (en la siguiente sección se demuestra).

1.2.4. Estudio de la varianza residual

Un estimador insesgado de σ^2 es la varianza residual S_R^2 , es decir, la suma de los residuos al cuadrado, corregida por los grados de libertad apropiados (en este caso $n - \dim(\mathcal{V})$):

$$S_R^2 = \frac{\sum e_i^2}{n-(k+1)} = \frac{\|e\|^2}{n-k-1} = \frac{\left\|Y - \hat{Y}\right\|^2}{n-k-1}$$

Si nos fijamos en que:

$$\left\|Y - \hat{Y}\right\|^2 = Y'(I-H)(I-H)Y = Y'(I-H)Y$$

y sabiendo que la matriz $I - H$ es simétrica e idempotente, si recordamos la proposición 1.12 y demostramos $\mu(I-H)\mu' = 0$, podemos determinar que la distribución de S_R^2:

Dado que $\mu \in \mathcal{V}$ y sabiendo que $\mu = XB$:

$$(I-H)\mu = (I-H)XB = 0$$

Ya que el vector $(I - H)$ proyecta sobre $\mathcal{V}^{\perp}$.

Así llegamos a que:

$$\frac{\left\|Y - \hat{Y}\right\|^2}{\sigma^2} = \frac{Y'(I-H)Y}{\sigma^2} =$$

$$\boxed{\frac{(n-k-1)S_R^2}{\sigma^2} \equiv \chi^2_{n-k-1}} \tag{1.2}$$

Además si tomamos esperanzas en ambos lados de la igualdad obtenemos:

$$\frac{(n-k-1)\cdot \mathbb{E}(S_R^2)}{\sigma^2} = n-k-1$$

$$\boxed{\mathbb{E}(S_R^2) = \sigma^2} \tag{1.3}$$

Lo siguiente que haremos es mirar si $\hat{\beta}$ y S_R^2 **son independientes**. Esto es cierto dado que se cumple que $(X'X)^{-1}X' \cdot (I - H) = 0$:

$$(X'X)^{-1}X' \cdot (I-H) = (X'X)^{-1}X' - (X'X)^{-1}X' \cdot X(X'X)^{-1}X' = 0$$

Observación: El lema de Fisher 1.13 no es más que el resultado de aplicar los resultados vistos en esta sección al caso particular del modelo:

$$y_i = \beta_0 + \varepsilon_i \iff X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

En este caso tenemos que $\dim(V) = 1$

1.2.5. Distribución de $\hat{\beta}$ y contrastes de hipótesis individuales sobre los coeficientes $\hat{\beta}_i$

$$\mathbb{E}\left(\hat{\beta}\right) = (X'X)^{-1}X' \underbrace{X\beta}_{\mathbb{E}(Y)} = \beta$$

$$\mathbb{V}\left(\hat{\beta}\right) = \sigma^2 I_n \cdot (X'X)^{-1}X' \cdot X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

Como $\hat{\beta}$ es una combinación lineal de Y (que sigue una distribución normal):

$$\hat{\beta} \equiv N_{k+1}\left(\beta, \sigma^2(X'X)^{-1}\right)$$

Y la regresión simple, es un caso particular de esta fórmula.

Notação: $q_{ij} \equiv$ entrada i, j de la matriz $(X'X)^{-1}$

Consecuencias:

- ¿Cuál es la distribución marginal de $\hat{\beta}_j$ a partir de la que hemos visto de la conjunta? Como vimos en el tema 1, es también una normal, con el correspondiente valor del vector β como media y el elemento j, j de la diagonal.

$$\hat{\beta}_j \equiv N\left(\beta_j, \sigma^2 q_{jj}\right)$$

Ahora, podemos estandarizar:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{q_{jj}}} \equiv N(0,1)$$

Y utilizando que S_R es independiente de σ y la definición de $t-$student tenemos:

$$\frac{\hat{\beta}_j - \beta_j}{S_R\sqrt{q_{jj}}} \equiv t_{n-k-1}$$

¿Cuál es el intervalo de confianza?

$$IC_{1-\alpha}(\beta_j) \equiv \left[\hat{\beta}_j \mp t_{n-k-1;\frac{\alpha}{2}} \underbrace{S_R\sqrt{q_{jj}}}_{\text{Error típico de } \beta_j}\right]$$

Y, como en regresión simple, estudiamos $H_0 : \beta_j = 0$:

$$R = \left\{\frac{\left|\hat{\beta}_j\right|}{S_R\sqrt{q_{jj}}} > t_{n-k-1;\frac{\alpha}{2}}\right\}$$

En las traspas encontramos una salida de regresión múltiple de R. La columna *estimate* es el vector $\hat{\beta}$, el p-valor es del contraste $H_0 : \beta_j = 0$.

1.2.6. Combinaciones lineales de los coeficientes

Vamos a querer constrastar cosas del estilo ¿las 2 variables influyen igual?

Para ello, transformamos eso en $H_0 : \beta_1 = \beta_2 \to H_0 : \beta_1 - \beta_2 = 0$, entonces tenemos una combinación lineal $a \in \mathbb{R}^{k+1}$, tal que $H_0 : a'\hat{\beta} = 0$

Para poder hacer esto, lo primero ha sido estimar $\hat{\beta}$ y lo siguiente será saber su distribución.

$$a'\hat{\beta} = N\left(a'\beta, \underbrace{\sigma^2 a'(X'X)^{-1}a}_{\left(E.T.(a'\hat{\beta})\right)^2}\right) \to \frac{a'\hat{\beta} - a'\beta}{S_R\sqrt{a'(X'X)^{-1}a}} \equiv t_{n-k-1}$$

Y con esto, ya podemos construir el intervalo de confianza para una combinación lineal de los parámetros:

$$IC_{1-\alpha}(a'\beta) = \left[a'\hat{\beta} \mp t_{n-k-1;\frac{\alpha}{2}} E.T.(a'\hat{\beta})\right]$$

La región de rechazo correspondiente es:

$$R = \left\{ \frac{|a'\hat{\beta}|}{S_R\sqrt{\sigma^2 a'(X'X)^{-1}a}} > t_{n-k-1;\frac{\alpha}{2}} \right\} = \left\{ \frac{|a'\hat{\beta}|}{E.T.(\hat{\beta})} > t_{n-k-1;\frac{\alpha}{2}} \right\}$$

Ejercicio: ¿Y si queremos hacer un contraste unilateral $a'\beta > 0$?

Hecho por Dejuan. Se aceptan correcciones.

Antes hemos contrastado que $a'\beta$ se queda dentro del intervalo de confianza. La región de rechazo será fuera del intervalo de confianza. El razonamiento es muy similar al caso de regresión simple (1.1.7)

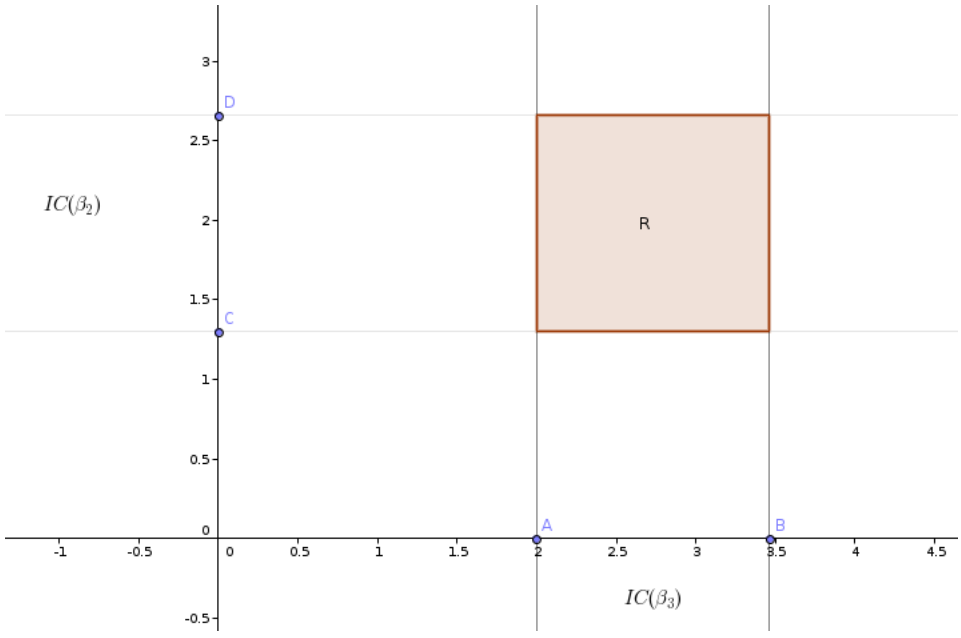
Como podemos salirnos por arriba o por abajo, aparece el valor absoluto. Si queremos que el **contraste sea unilateral**, rechazaremos cuando $a'\beta$ (corregido por su error típico) sea demasiado negativo. Esto es

$$R = \left\{ \frac{a'\hat{\beta}}{S_R\sqrt{\sigma^2 a'(X'X)^{-1}a}} < -t_{n-k-1;\alpha} \right\} = \left\{ \frac{a'\hat{\beta}}{E.T.(\hat{\beta})} < -t_{n-k-1;\alpha} \right\}$$

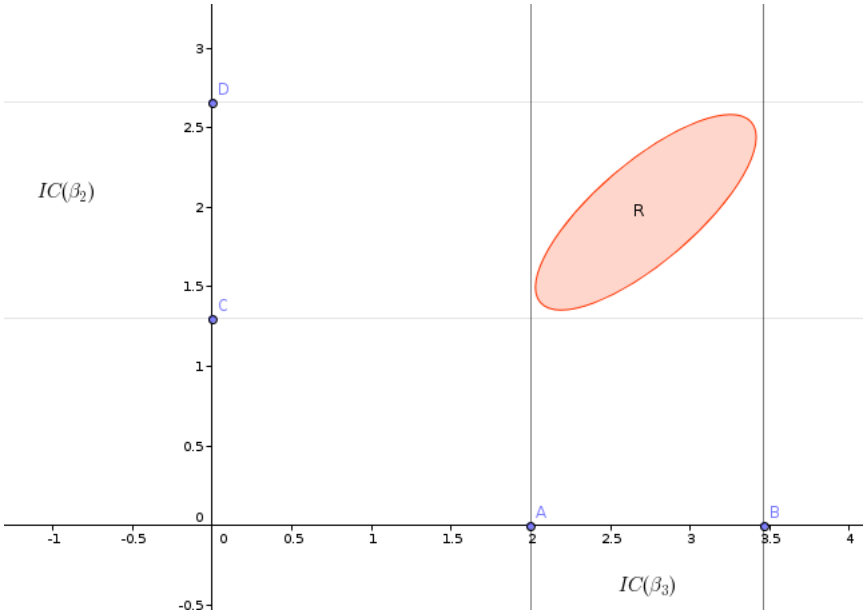
1.2.7. Inferencia sobre subconjuntos de coeficientes

Todos los contrastes hechos hasta ahora son muy fáciles porque se basan en una única normal. Nuestros contrastes han sido del tipo $\beta_i = 0$. En esta sección vamos a estudiar contrastes del tipo $H_0 : \beta_1 = 0, \beta_3 = 0$.

La primera aproximación puede ser calcular la región de confianza de β_1 y la de β_3 y tomar la intersección. Es decir:



Pero al no ser independientes, no estamos teniendo en cuanto las correlaciones, la información que me da saber β_1 para saber β_3.



Vamos a ello formalmente: Sea $\beta_{(p)}$ un conjunto de coeficientes de β. Sea $\hat{\beta}_{(p)}$ el correspondiente subconjunto de estimadores.

Sea Q_q la submatriz de $(X'X)^{-1}$ cuyas filas y columnas corresponden a las coordenadas del vector $\beta_{(p)}$.

$$\hat{\beta}_{(p)} \equiv N_p\left(\beta_{(p)}, \sigma^2 Q_p\right)$$

Si este es nuestro estimador, vamos a estandarizarlo (utilizando propiedades del tema 1).

$$\frac{(\hat{\beta}_{(p)} - \beta_{(p)})' Q_p^{-1} (\hat{\beta}_{(p)} - \beta_{(p)})}{\sigma^2} \equiv \chi_p^2$$

Tenemos el problema de que no conocemos σ y tenemos que estimarlo. Para ello, vamos a seguir un proceso parecido a 1.1.6. Para ello necesitamos:

Distribución $F_{n,m}$

Definición 1.7 **Distribución** $F_{n,m}$.

$$F_{n,m} \equiv \frac{\chi_m^2/m}{\chi_n^2/n}$$

Es muy habitual que aparezca la F al comparar varianzas.

Sabiendo lo que es una F, vamos a estudiar qué ocurre al cambiar σ por S_R

$$\frac{(\hat{\beta}_{(p)} - \beta_{(p)})' Q_p^{-1} (\hat{\beta}_{(p)} - \beta_{(p)})}{S_R^2} = \frac{\frac{(\hat{\beta}_{(p)} - \beta_{(p)})' Q_p^{-1} (\hat{\beta}_{(p)} - \beta_{(p)})}{\sigma^2}}{\frac{S_R^2}{\sigma^2}}$$

En el numerador tenemos una χ^2_p, pero nos faltaría dividir por los grados de libertad para tener una F, entonces:

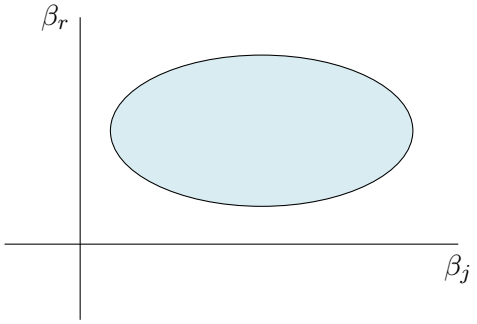
$$\frac{\frac{(\hat{\beta}_{(p)}-\beta_{(p)})'Q_p^{-1}(\hat{\beta}_{(p)}-\beta_{(p)})}{p\sigma^2}}{\frac{S_R^2}{\sigma^2}} = \frac{\chi^2_p/p}{\chi^2_{n-k-1}/n-k-1} \equiv F_{p,n-k-1}$$

Conclusión

$$\frac{(\hat{\beta}_{(p)}-\beta_{(p)})'Q_p^{-1}(\hat{\beta}_{(p)}-\beta_{(p)})}{pS_R^2} \equiv F_{p,n-k-1}$$

$$1-\alpha = P\left((\hat{\beta}_{(p)}-\beta_{(p)})'\left(S_R^2Q_p\right)^{-1}(\hat{\beta}_{(p)}-\beta_{(p)}) \leq pF_{p,n-k-1}\right)$$

Esto nos da un elipsoide de confianza, es decir, sabemos que caerá en la región del elipsoide con probabilidad $1-\alpha$:



Ejemplo: *Vamos a querer contrastar si son iguales los coeficientes β_2, β_3 las variables "Income" y "Miles". La hipótesis es: $H_0 : \beta_2 = \beta_3$ a nivel $\alpha = 0.05$*

Aquí damos los datos necesarios para completar (en la realidad, los obtendremos de R:

$$S_R^2Q_{(2)} = \begin{pmatrix} 259.45 & 7.89 \\ 7.89 & 2.52 \end{pmatrix}$$

Vamos a proceder al contraste. Construimos el estadístico para construir la región de rechazo:

$$t = \frac{|\hat{\beta}_2 - \hat{\beta}_3|}{E.T.(\hat{\beta}_2 - \hat{\beta}_3)} = \frac{1.725}{15.687} \not\geq t_{45;0.025}$$

Por tanto no podemos rechazar la hipótesis nula de que $\beta_2 = \beta_3$.

El error típico se ha calculado es:

$$\sqrt{\mathbb{V}\left(\hat{\beta}_2 - \hat{\beta}_3\right)} = \sqrt{(1,-1)S_R^2Q_p(1,-1)} = 15.687$$

Y los 45 grados de libertad los obtenemos de R.

1.2.8. Predicción

En el caso de regresión múltiple querremos predecir valores medios o exactos de Y_0 cuando conocemos los valores exactos de las variables regresoras $X_0 = (1, x_{01}, \ldots, x_{0k})'$.

Confianza de $Y_0|X_0$

$$m_0 = E(Y_0|X_0) = \beta' X_0 \to \hat{m}_0 \equiv N\left(\beta' X_0, \sigma^2 X_0'(X'X)^{-1}X_0\right) \tag{1.4}$$

$$\hat{m}_0 = \hat{\beta} X_0$$

Ahora ya podemos calcular el intervalo de confianza para Y_0 dado un vector X_0.

$$IC_{1-\alpha}(m_0) = \left[\hat{m}_0 \mp t_{n-k-1;\frac{\alpha}{2}} S_R \sqrt{X_0'(X'X)^{-1}X_0}\right]$$

Predicción de Y_0

$$IC_{1-\alpha}(Y_0) = \left[\hat{Y}_0 \mp t_{n-k-1;\frac{\alpha}{2}} S_R \sqrt{1 + X_0'(X'X)^{-1}X_0}\right]$$

2. Análisis de la varianza (ANOVA)

2.1. Conceptos previos

Las variabilidades se miden mediante la suma de cosas al cuadrado. Vamos a definir 3 sumas de cuadrados:

Suma de cuadrados total

Definición 2.1 **Suma de cuadrados total**. *Medimos la variabilidad total con:*

$$SCT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Esta suma de cuadrados, mide la variabilidad "real" de los datos. Aquí no hay ningún modelo.

Suma de cuadrados de la regresión

Definición 2.2 **Suma de cuadrados de la regresión**. *Medimos la variabilidad explicada por el modelo con:*

$$SCR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

Esta suma de cuadrados, mide la variabilidad según el modelo. En caso de $SCT = SCR$, tendríamos que el modelo es perfecto.

Suma de cuadrados total

Definición 2.3 **Suma de cuadrados total**. *Medimos la variabilidad no explicada.*

$$SCE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Intuitivamente, si el modelo estuviera bien construido sería razonable esperar que $SCT = SCR + SCE$. Vamos a complicarnos la vida y utilizar la interpretación geométrica para ver que esa relación es cierta:

SCT vectorialmente

$$SCT = ||Y - \overline{Y}1_n||^2$$

Vamos a ver una manera complicada de entender la media muestral: Proyección de un vector al espacio de proyección de los vectores con las mismas coordenadas. ¿Esto a qué viene?

$$\overline{Y}1_n = \frac{1}{n}1_n 1_n' Y = MY$$

Siendo

$$M = \begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

Entonces, podemos construir:

$$SCT = ||Y - \overline{Y}1_n||^2 = ||Y - MY||^2 = ||(I - M)Y||^2 = Y'(I - M)Y$$

SCR vectorialmente

$$SCR = ||HY - MY||^2 = Y'(H - M)'(H - M)Y$$

Necesitamos ver que $H - M$ es una matriz simétrica e idempotente. Recordamos que H es la matriz de proyección sobre $\mathcal{V}$.

$$(H - M)(H - M) = HH - MH - HM + MM \overset{(1)}{=} H - M$$

(1) sabemos que $HH = H$ y $MM = M$ porque ambas son idempotentes. Pero para tener una M restando, necesitaríamos $MH = HM = M$.

Geométricamente, es fácil ver que $HM = M$. Esto se debe a que M es la proyección sobre un espacio vectorial más pequeño e incluido en H, entonces al haber proyectado con M, volver a proyectar con H no nos cambia nada.

Por otro lado, $MH = M$. En este caso, primero proyectamos en el subespacio grande, pero como acabamos proyetando sobre el pequeño.

Ejercicio: Demostrar algebraicamente $MH = HM = M$.

SCE vectorialmente

$$SCE = ||HY - H||^2 = Y'(I - H)Y$$

Proposición 2.1 (Relación de las sumas de cuadrados)**.**

$$SCT = SCR + SCE$$

Demostración.

$SCT = Y'(I-M)Y = Y'(H-M+I-H)Y = Y'(H-M)Y+Y'(I-H)Y = SCR+SCE$

□

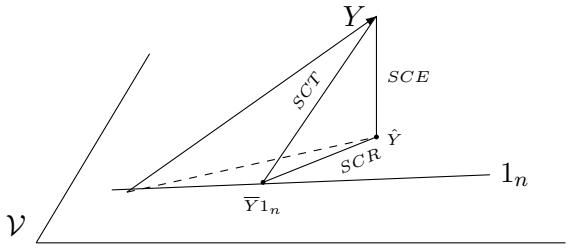


Figura II.1: *Representación de los vectores SCE, SCT, SCR sobre el espacio vectorial $\mathcal{V}$.*

2.1.1. Contraste de la regresión

Queremos contrastar $H_0 : \beta_1 = ... = \beta_n = 0$.

La idea intuitiva es que si SCR es muy pequeño, tenemos poca variabilidad explicada sería razonable que aceptáramos la hipótesis nula de que las variables regresoras no influyen mucho.

Por el contrario, si SCR es muy **grande**, tenemos mucha variabilidad explicada, por lo que la hipóteis nula no es muy razonable y deberíamos **rechazar**. Por el contrario, si SCR es **pequeña** aceptamos H_0.

Para ello, es imprescindible definir "grande" y "pequeño". Deberían tener las mismas unidades. No puede ser que por cambiar de unidades afecte a la validez de la hipótesis. Además, necesitamos saber la distribución de SCR bajo la hipótesis nula.

Proposición 2.2 (Distribución SCR en $H_0 : \forall i\ \beta_i = 0$)**.**

$$\frac{SCR}{\sigma^2} \equiv \chi^2_k$$

Demostración. Tenemos $SCR = Y'(H-M)Y$, una forma cuadrática, teniendo $H-M$ simétrica e idempotente (por construcción).

Ahora, para poder aplicar la proposición del tema 1, necesitamos $\mu'(H-M)\mu = 0$.

Sabemos que μ, bajo H_0 tiene todas sus coordenadas iguales, con lo que está en el subespacio 1_n, por lo que las proyecciones sobre V y sobre 1_n de μ serán μ. Esto quiere decir que:

$$H\mu = \mu \quad M\mu = \mu \rightarrow (H - M)\mu = 0$$

Por otro lado, los k grados de libertad:

$$gl = \mathrm{Rg}(H - M) = \mathrm{traza}(H - M) = \mathrm{traza}(H) - \mathrm{traza}(M) = (k + 1) - 1 = k$$

Esto se debe a que la traza de una matriz de proyección es la dimensión del espacio en el que proyectamos.

□

As usual, como no conocemos σ^2, en la práctica utilizamos:

Proposición 2.3.

$$\frac{SCR}{k \cdot S_R^2} = \frac{\frac{SCR}{k}}{\frac{SCE}{n-k-1}} = \frac{\chi_k^2}{\chi_{n-k-1}^2} \sim F_{k;n-k-1}$$

Demostración. Para ello necesitamos que SCR y SCE sean independientes. Son normas al cuadrado de vectores ortogonales. Estadísticamente, la ortogonalidad de vectores normales se traduce en independencia.

Para ello, vamos a ver que el producto de las matrices es 0 (utilizando la propiedad 1.12)

$$(H-M)(I-H) = H - H - M + \underbrace{MH}_{M} = 0$$

Sabemos que $MH = M$ por que, por un lado es proyectar en V después de haber proyectado en un subespacio de V (1_n) y en el otro, tenemos la proyección directamente sobre 1_n.

□

2.1.2. Tabla anova

ANOVA

Para este caso, vamos a construir la tabla **ANOVA** (análisis de la varianza)

Este es el ejemplo de una tabla ANOVA simple:

Fuente de variación	SC	gl	CM	F	p-valor
Explicada	SCR	k	$\frac{SCR}{k}$	$\frac{SCR/k}{SCE/(n-k-1)}$	...
No explicada	SCE	n-k-1	$\frac{SCE}{n-k-1}$	...	...
total					

2.1.3. Coeficiente de determinación

Coeficiente de determinación

Definición 2.4 **Coeficiente de determinación**. *Es una medida de la bondad del ajuste en el modelo de regresión múltiple*

$$R^2 = \frac{SCR}{SCT}$$

Propiedades:

- $R^2 \leq 1$ (ya que un cateto siempre es menor o igual que la hipotenusa)
- Vamos a ver los casos extremos.

  $R^2 = 1$ quiere decir que existe una relación lineal exacta debido a que

  $$R^2 = 1 \implies SCR = SCT \implies SCE = 0 \implies \forall i \; e_i = 0$$

  De hecho, podemos interpretar R^2 como un coeficiente de correlación múltiple entre las k variables X y la Y.

  $R^2 = 0$ quiere decir que no existe una relación lineal debido a que

  $R^2 = 0 \implies \Sigma(\hat{Y}_i - \overline{Y})^2 = 0 \implies \hat{Y}_i = \overline{Y} \implies$ ninguna X aporta información para calcular Y

- Se verifica que el estadístico $F = \dfrac{R^2}{1-R^2}\dfrac{n-k-1}{k}$

  Demostración. Esta identidad se obtiene inmediatamente utilizando la definición de R^2:

  $$\frac{R^2}{1-R^2}\frac{n-k-1}{k} = \frac{SCR}{SCT-SCR}\frac{n-k-1}{k} = \frac{SCR/k}{SCE/(n-k-1)} = F$$

  □

Este coeficiente es útil en el contraste que estudiábamos anteriormente $H_0 : \forall i \; \beta_i = 0$. En este contraste, es equivalente para rechazar o aceptar

- SCR es "grande".
- F es "grande".
- R^2 se aproxima a 1.

El coeficiente de determinación para comparar distintos modelos de regresión entre sí tiene el siguiente inconveniente: Siempre que se añade una nueva variable regresora al modelo, R^2 aumenta, aunque el efecto de la variable regresora sobre la respuesta no sea significativo. ¿Porqué aumenta R^2? Porque la dimensión de V aumenta provocando que disminuya SCE (por algo que me he perdido) y al disminuir SCR, aumenta SCR (porque su suma es constante) y aumenta R^2.

Por ello, definimos:

Coeficiente de determinación ajustado

Definición 2.5 **Coeficiente de determinación ajustado**.

$$\overline{R}^2 = r^2 = 1 - \frac{SCE/(n-k-1)}{SCT/(n-1)}$$

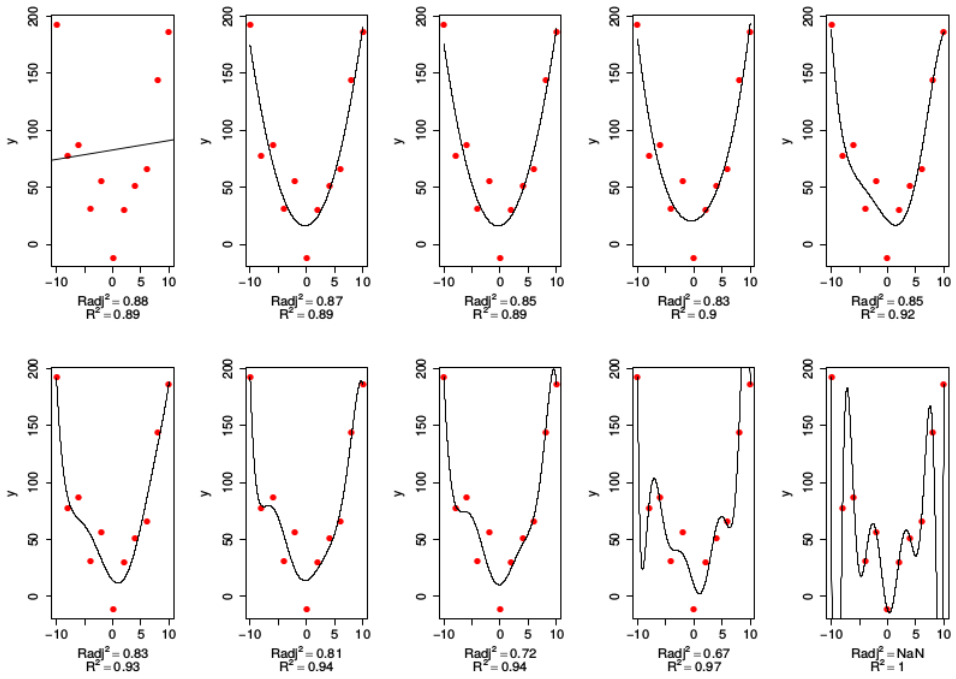
Esta nueva definición penaliza un aumento de las variables utilizadas. De esta manera, al aumentar K, sabemos que R^2 aumenta pero $\overline{R}^2$ puede aumentar o no. Depende de si la ganancia de información es mayor que la penalización de añadir más variables.

Vamos a ver un ejemplo realmente ilustrativo.

Ejemplo: *Vamos a utilizar el modelo de regresión múltiple para ajustar el grado del polinomio.*

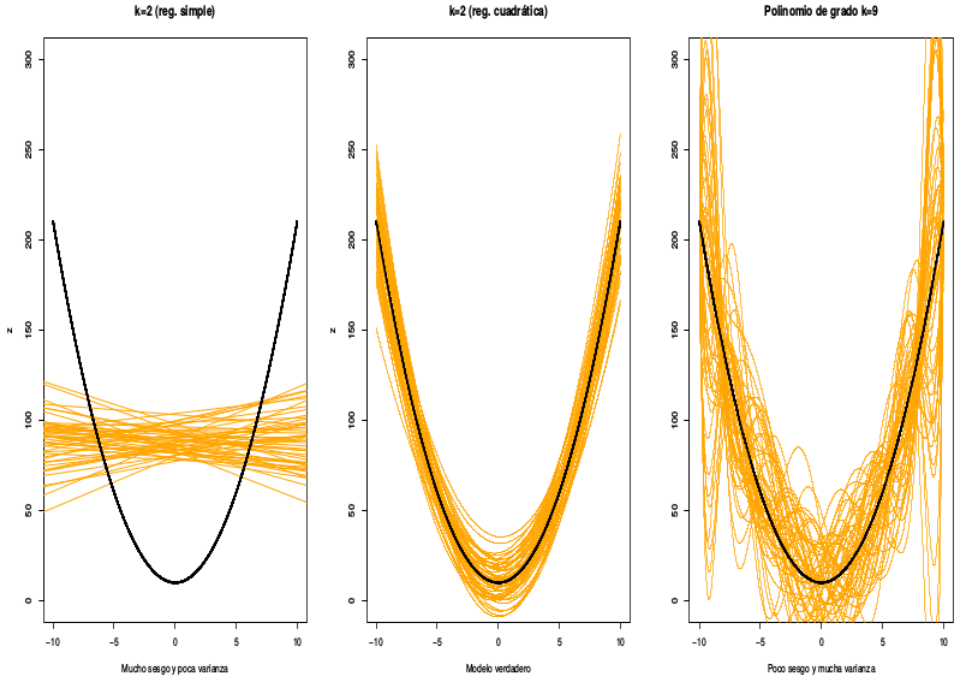
El proceso a seguir podría ser: voy calculando R^2 para todos los grados que vaya teniendo hasta que tenga $R^2 = 1$, que hace que podamos explicar toda la varianza con estos datos.

Al aumentar el grado, obtenemos polinomios cada vez mejores (para estos datos).



Si tomáramos una nueva muestra de datos, el modelo que tiene $R^2 = 1$ sería una mierda muy probablemente. En este caso, sería una bazofia absoluta ya que los datos han sido generados parabólicamente. Es por ello que el modelo con mayor $\overline{R}^2$ es el segundo.

Ejemplo: *Si un modelo es demasiado simple, tendremos muy poca varianza pero mucho sesgo. Por el otro lado, si el modelo es demasiado complejo, tenemos mucha varianza pero muy poco sesgo. Vamos a verlo con un ejemplo:*



En este ejemplo, el primer caso, demasiado simple tiene muy poca varianza. Todas las pendientes son similares, pero por el contrario, siempre estimamos mal.

En el tercer caso, si cogemos un único polinomio, vemos que es un poco malo, pero si tomáramos la media de todos ellos, más o menos se asemeja bastante a la parábola, es decir, la esperanza es el valor esperado con lo que el sesgo sería muy pequeño. Por el contrario, cada polinomio es de su padre y de su madre, es decir, tiene mucha varianza estimar así.

Ejemplo: *Vamos a calcular el coeficiente de determinación para el caso de regresión simple:*

$$R^2 = \frac{SCR}{SCT} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})} = \frac{\sum(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y})^2}{S_{yy}}$$
$$= \frac{\sum(\overline{y} + \hat{\beta}_1(x_i - \overline{x}) - \overline{y})^2}{S_{yy}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = r^2$$

Coeficiente de correlación

Definición 2.6 ***Coeficiente de correlación****.*

Dicha r se conoce como coeficiente de correlación y aparece en una de las múltiples expresiones del coeficiente $\hat{\beta}_1$ de la regresión: $\hat{\beta}_1 = r \cdot \frac{S_y}{S_x} \implies \hat{\beta}_1^2 = r^2 \cdot \frac{S_{yy}}{S_{xx}}$. Expandiendo la expresión se obtiene:

$$r = r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

2.2. Contrastes de hipótesis lineales

Queremos contrastar $H_0 : A\beta = 0$, donde A es una matriz $p \times (k+1)$ con $\text{Rg}(A) = p < k+1$.

Ejemplo: *En el modelo $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ queremos contrastar*

$$H_0 : \beta_1 = \beta_2 \; \beta_0 = 0 \iff A\beta = 0 \to A = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Modelo reducido (M_0)

Definición 2.7 **Modelo reducido (M_0).** *Es el modelo resultante al imponer las restricciones de H_0.*

Modelo completo

Ejemplo: *Tomando el **Modelo completo** del ejemplo anterior*

$$\left.\begin{array}{c} Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \\ H_0 : \beta_1 = \beta_2 \; \beta_0 = 0 \end{array}\right\} \to Y_i = \beta_1(x_{i1} + x_{i2}) + \beta_3 x_3 + \varepsilon_i$$

El modelo simple es $Y_i = \beta_1(x_{i1} + x_{i2}) + \beta_3 x_3 + \varepsilon_i$

Cómo elegir un modelo Nos podemos plantear qué modelo podemos elegir. La idea sería elegir siempre lo simple, pero tenemos que ver cuánto perdemos con la simplificación. A esto se le denomina **Principio de incremento relativo de la variabilidad**

Principio de incremento relativo de la variabilidad

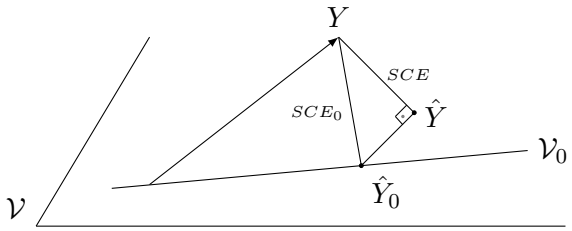


Figura II.2: *interpretação geométrica de erro inexplicável no modelo completo e reduzido, em que $\mathcal{V}_0$ é o espaço vectorial em que as estimativas são obtidas modelo reduzida (do exemplo $dim(\mathcal{V}_0) = 2$, $dim(\mathcal{V}) = 4$).*

La idea es que $SCE_0 > SCE$. Geométricamente, vemos que SCE_0 es la hipotenusa y SCE es un cateto. Por otro lado, la variabilidad no explicada de un modelo simple siempre va a ser menor que la SCE de un modelo complejo para los mismos datos.

Entonces la idea sería rechazar H_0 cuando se pierda mucho al considerar M_0 en lugar de M, es decir cuando $SCE_0 - SCE$ sea muy grande. Para evitar problemas generados por cambios de escala, tomamos

$$\frac{SCE_0 - SCE}{SCE}$$

¿Y esto que distribución tiene? ¿Cómo podemos definir ese "suficientemente grande"? Las sumas de cuadrados (SCE) son χ^2 y el cociente de χ^2's es una F. Vamos a estudiarlo formalmente:

Distribución de $(SCE_0 - SCE)/SCE$ Sea X la matriz de diseño del modelo reducido correspondiente. "Hat matrix" es $H_0 = X_0(X_0'X_0)^{-1}X_0'$, donde tenemos $\mathrm{Rg}(H_0) = \mathrm{traza}(H_0) = k + 1 - p$. El rango es la traza porque es una matriz de proyección y eso es lo que debería ser debido a .....

Vamos a construir el estadístico:

$$SCE_0 - SCE = Y'(I - H_0)Y - Y'(I - H)Y = Y'(H - H_0)Y$$

Vamos a ver si $H - H_0$ es simétrica e idempotente, teniendo que $H_0H_0 = H_0$, por ser una matriz de proyección.

$$(H - H_0)(H - H_0) = H - H_0H - HH_0 + H_0 = H_0$$

Aquí tenemos que $H_0H = HH_0 = H_0$ por el mismo argumento que en 2.1.1 Ambas son matrices de proyección sobre subespacios en los que uno está contenido dentro del otro, por lo que da igual el orden en el que proyectar y además, será lo mismo que proyectar sobre el pequeño.

Ahora, necesitamos ver que $\mu'(H - H_0)\mu = 0$ para poder tener que $SCE_0 - SCE \sim \chi^2$. Aquí el razonamiento vuelve a ser muy parecido.

$$(H - H_0)\mu = H\mu - H_0\mu = \mu - \mu = 0$$

Proyectar μ con H y con H_0 no cambia μ, debido a que μ está en el subespacio de proyección pequeño V_0. ¿Porqué sabemos que $\mu \in V_0$? Porque, asumiendo H_0

$$\frac{SCE_0 - SCE}{\sigma^2} \overset{H_0}{\equiv} \chi^2_p \qquad \frac{SCE}{\sigma^2} \equiv \chi^2_{n-k-1}$$

Si fueran independientes, su cociente sería una F. Vamos a ver la independencia utilizando 1.12 y utilizando que $HH = H$ y $HH_0 = H_0H = H_0$ por ser matrices de proyección.

$$(H - H_0)(I - H) = H - H_0 - H + H_0 = 0 \rightarrow \text{ independientes}$$

Concluimos que

$$\frac{(SCE_0 - SCE)/p}{SCE/(n - k - 1)} \overset{H_0}{\equiv} F_{p;n-k-1}$$

Cuya región de rechazo será para un nivel α será

$$R_\alpha = \left\{ \frac{(SCE_0 - SCE)/p}{SCE/(n - k - 1)} > F_{p,n-k-1;\alpha} \right\}$$

Observación: El contraste $H_0 : \beta_1 = ... = \beta_k = 0$ es un caso particular de esto que hemos visto, tomando en $A\beta = 0$ la matriz A como:

$$A = \begin{pmatrix} 0 & \\ \vdots & I_k \\ 0 & \\ \underbrace{\qquad}_{(1)} & \end{pmatrix}$$

(1): por β_0.

En este caso, el modelo reducido sería $Y_i = \beta_0 + \varepsilon_i$, con lo que

$$SCE_0 = \sum (Y_i - \hat{Y}_i)^2 \overset{(2)}{=} \sum (Y_i - \overline{Y})^2 = SCT$$

(2):debido a que $Y_1, ..., Y_n \sim N(\beta_0, \sigma^2)$, con lo que $\hat{\beta}_0 = \overline{Y}$.

Como $SCE_0 = SCT$, tenemos lo que ya habíamos obtenido:

$$SCE_0 = SCT \rightarrow SCE_0 - SCE = SCT - SCE = SCR$$

2.2.1. Explicación de la tabla ANOVA

R siempre está comparando 2 modelos, el reducido y el completo.Vamos a entenderla por inducción matemática:

En la primera fila, el modelo completo es el que incluye la primera variable y el modelo reducido es el que no incluye ninguna.

El la fila n, el modelo completo es el que incluye las n variables, y el modelo reducido es el que incluye las $n-1$ variables.

	Df	Sum Sq	Mean Sq	F value
x1	1	SC_1	SC_1	SC_1/MCE
x2	1	SC_{12}	SC_{12}	SC_{12}/MCE
x3	1	SC_{123}	SC_{123}	SC_{123}/MCE
Residuals	$n-k-1$	SCE	$MCE = SCE/(n-k-1)$	

Es importante darse cuenta que los números de la columna $SumSq$ dependen del orden de las variables, ya que no es lo mismo pasar de no tener variables a tener la primera variable (que puede no influir nada) a tener la segunda (que puede ser tremendamente explicativa).

Sin embargo, la suma es **constante**. Vamos a verlo

Demostración.

$$\begin{aligned} SC_1 &+ SC_{12} &+ SC_{123} &= \\ (SCE_0 - SCE_1) &+ (SCE_1 - SCE_{12}) &+ (SCE_{12} - SCE_{123}) &= \end{aligned}$$

$$SCE_0 - SCE_{123} = SCT - SCE = SCR$$

□

Ejemplo: *Vamos a ver una tabla anova normal, con los datos del consumo de combustible en EEUU:*

```
reg <- lm(FuelC ~ Drivers+Income+Miles+MPC+Tax, data=fuel2001)
anova(reg)

Response: FuelC
```

```
            Df     Sum Sq    Mean Sq   F value     Pr(>F)
Drivers      1 3.5301e+14 3.5301e+14 2273.2167 < 2.2e-16 ***
Income       1 6.7563e+11 6.7563e+11    4.3507  0.0426945 *
Miles        1 2.1698e+12 2.1698e+12   13.9723  0.0005216 ***
MPC          1 5.2927e+11 5.2927e+11    3.4082  0.0714577 .
Tax          1 4.1208e+11 4.1208e+11    2.6536  0.1102978
Residuals 45 6.9882e+12 1.5529e+11
---
```

Vemos que el $p-valor$ de la última variable es 0.1102978. *Si vamos al análisis de regresión múltiple hecho al principio de la sección (1.2) vemos que el p-valor de Tax (nuestra última variable) es* 0.110298. *¿Casualidad o causalidad?*

La última columna de la fila n es el p-valor del contraste $\beta_n = 0$, tomando el modelo que tiene las $\beta_0, ..., \beta_{n-1}$. En el caso de la última fila, tenemos el p-valor de si $\beta_{ultimo} = 0$ en el modelo que incluye todas las demás variables. Este contraste es equivalente al que hacíamos en regresión múltiple para contrastar si $\beta_j = 0$. Podemos comprobar que los p-valores coinciden (0.11).

Ejemplo: *Vamos a ver ahora el contraste de si podemos utilizar el modelo simple frente al complejo.*

```

regsimple <- lm(FuelC ~ Drivers,data=fuel2001)
anova(regsimple, reg)

Model 1: FuelC ~ Drivers
Model 2: FuelC ~ Drivers + Income + Miles + MPC + Tax
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     49 1.0775e+13
2     45 6.9882e+12  4 3.7868e+12 6.0962 0.0005231 ***
---
```

Es de recibo comentar que: $RSS = SCE$

Por otro lado, son 4 grados de libertad porque imponemos 4 restricciones del modelo compuesto para el modelo simple (4 coeficientes nulos 0).

El p-valor obtenido, al ser tan pequeño nos dice que la ganancia de información es suficientemente grande como para tener que rechazar el modelo simple.

2.2.2. Análisis de influencia

En esta subsección vamos a estudiar cuánto de influyente puede ser una observación. En el caso de regresión simple, teníamos que algunos datos atípicos podían desviar mucho la recta (ver 1.1.2). ¿Qué ocurre en regresión múltiple? Aquí no podemos hacer razonamientos geométricos para entenderlo, pero es un tema importante que tratar.

Vamos a estudiarlo utilizando un ejemplo en el que queremos saber la cantidad de cierto medicamento en el hígado de una rata, tras recibir una dosis oral. La dosis recibida fue de 40 mg por kg de peso corporal. Tras cierto tiempo se sacrifican las ratas y se mide la cantidad de medicamento en su hígado.

Tenemos 19 observaciones, y tres variables regresoras (peso de la rata, peso del hígado y dosis recibida) para la variable respuesta (Proporción de dosis en el hígado)

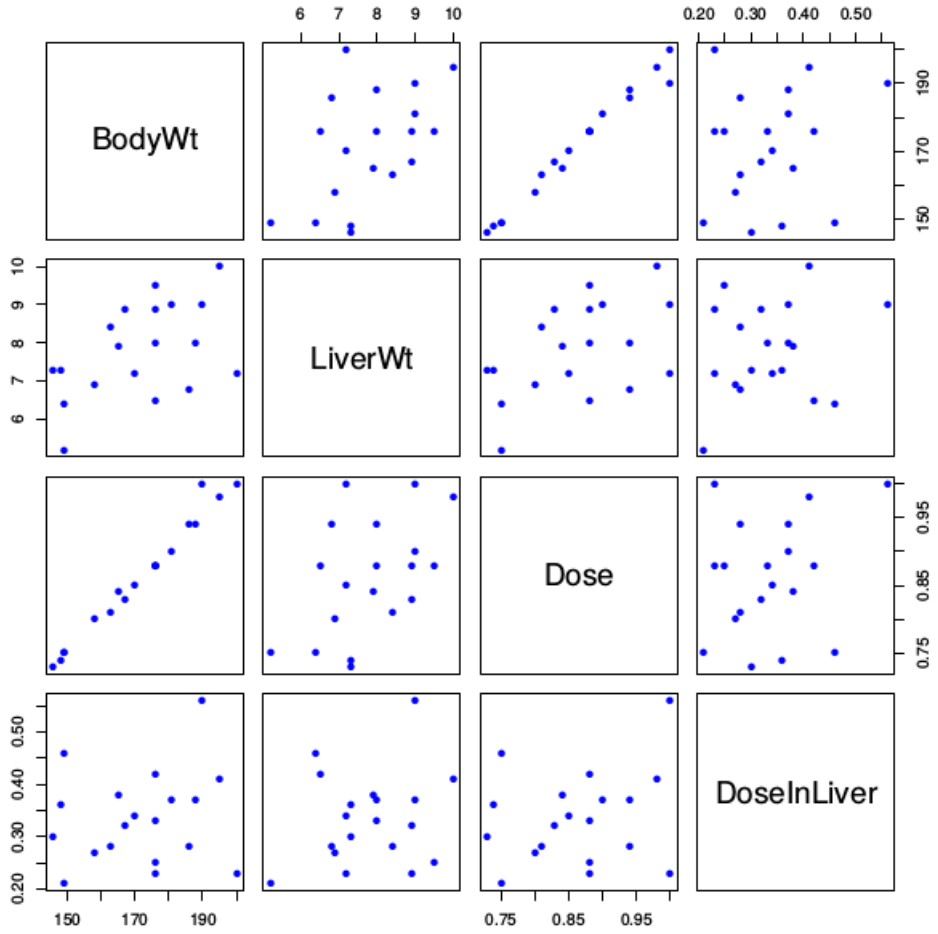
Vamos a ver contrastando con regresión simple cada $\beta_i = 0$. Estos son los resultados obtenidos:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1962346	0.2215825	0.886	0.388
BodyWt	0.0008105	0.0012862	0.630	0.537

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.22037	0.13573	1.624	0.123
LiverWt	0.01471	0.01718	0.856	0.404

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1330	0.2109	0.631	0.537
Dose	0.2346	0.2435	0.963	0.349

Vemos que los $p - valores$ son muy grandes, con lo que no podemos rechazar $\beta_i = 0$ y no debería haber relación entre la variable respuesta a partir de la regresora. Además, podemos corroborarlo gráficamente:



Aparentemente no hay relación entre la respuesta y las regresoras. **¿Y haciendo**

regresión múltiple? Tal vez contrastando $\beta_1 = ... = \beta_3 = 0$ obtengamos otra cosa (aunque sería de esperar que no).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.265922	0.194585	1.367	0.1919
BodyWt	-0.021246	0.007974	-2.664	0.0177
LiverWt	0.014298	0.017217	0.830	0.4193
Dose	4.178111	1.522625	2.744	0.0151

Hay unos cuantos valores extraños. ¿Ahora resulta que $BodyWt$ y $Dose$ (que además son prácticamente la misma variable, ya que existe una gran correlación entre ellas) si son influyentes? Pero si antes no lo eran... Además, el término independiente de $Dose$ es tremendamente distinto al anterior[6]. ¿Cómo puede ser esto posible? **La paradoja se produce por un dato atípico.** Vamos a verlo, pero para ello necesitamos un par de medidas, para poder entender lo atípico del dato.

Potencial en un punto
Leverage

Definición 2.8 **Potencial en un punto**. *El potencial (**leverage**) de un punto es el correpondiente h_i de la matriz de diseño X.*

Los potenciales definen las varianzas, ya que

$$\mathbb{V}(e_i) = \sigma^2(1 - h_i)$$

Demostración. Esto se debe a la distribución de los residuos.

Recordamos $e = Y - \hat{Y} = (I - H)Y$. Como $Y \equiv N(X\beta, \sigma^2 I_n)$, tenemos:

$$e \equiv N\left(0, (I - H)\sigma^2 I_n (I - H)'\right) \equiv N(0, \sigma^2(I - H))$$

□

Además, este potencial está tremendamente relacionado con la distancia de Mahalanobis d_M:

$$h_i = \frac{1}{n} + \frac{1}{n-1}\mathrm{d}_M^2(x_i, \overline{x})$$

Este potencial mide numéricamente la influencia de un único punto. Para valores > 0.5, el libro (sin justificación teórica) sugiere preocuparse porque tal vez sea un dato atípico. Otra manera (más natural) de medir la influencia, es recalcular el modelo eliminando el dato. Si cambia mucho, el dato es influyente. Si no cambia nada, el dato no es influyente. Esta es la idea que hay detrás de la **Distancia de Cook**.

Distancia de Cook

Definición 2.9 **Distancia de Cook**. *La distancia de Cook mide cómo cambia el vector de estimadores $\hat{\beta}$ cuando se elimina cada observación.*

Para ello, se utiliza la distancia de Mahalanobis (estandarizada) entre $\hat{\beta}, \hat{\beta}_i$.

Recordando que $\hat{\beta} \equiv N_{k+1}(\beta, \sigma^2(X'X)^{-1})$ y que la matriz de covarianzas se puede estimar como $\hat{\beta} = S_R^2(X'X)^{-1}$, entonces:

[6] **Se deja como ejercicio para profundizar:** *Demostrar que los β_i de un modelo de regresión simple son iguales a los β_i de cada regresión simple por cada variable regresora si éstas son independientes (o incorreladas).*

$$D_i = \underset{\textit{estandarizada}}{d_M^2(\hat{\beta}, \hat{\beta}_{(i)})} = \frac{[\hat{\beta} - \hat{\beta}_{(i)}]'(X'X)[\hat{\beta} - \hat{\beta}_{(i)}]}{(k+1)S_R^2}$$

Propiedades

- ¿Cuándo es grande esta distancia? Lo que se hace es calibrarlo comparando con $F_{k+1,n-k-1;\alpha}$. En general $D_i > 1$ suele ser relevante.
- Se puede calcular de una manera más simple utilizando los valores ajustados:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j(i) - \hat{Y}_j\right)^2}{(k+1)S_R^2}$$

- Y también está relacionado con el potencial y los residuos.

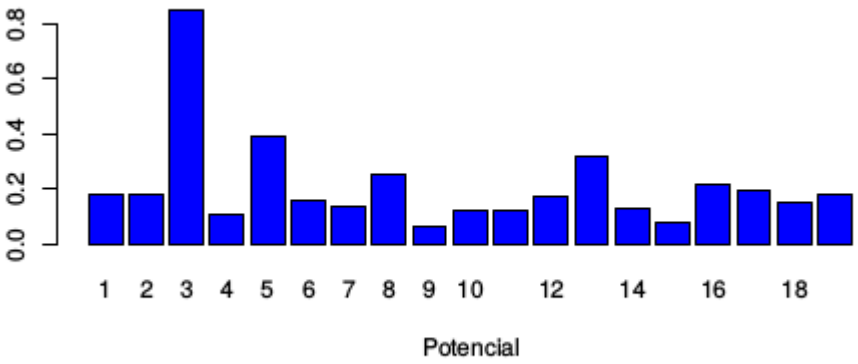
$$D_i = \frac{1}{k+1} r_i^2 \frac{h_i}{1-h_i}$$

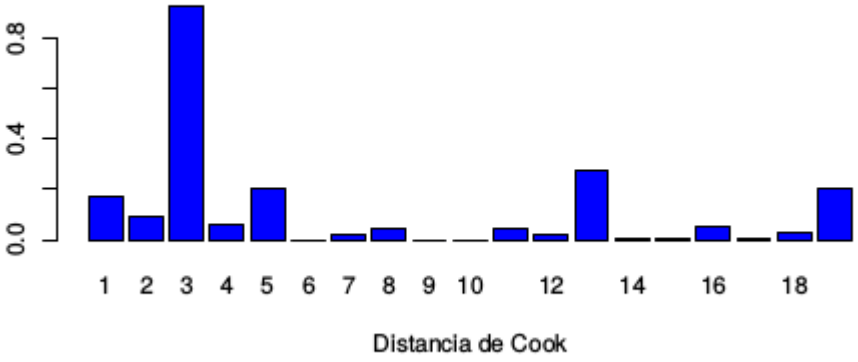
  donde $r_i = e_i/(S_R\sqrt{1-h_i})$, los residuos estandarizados.

En el ejemplo (1.1.2), tanto el punto rojo como el verde tienen un alto potencial, pero el rojo tiene una distancia de Cook mucho menor, ya que no es muy atípico (en tanto en cuanto el modelo se ajusta bien a ese dato). En cambio, el punto verde tiene una distancia de Cook mucho mayor.

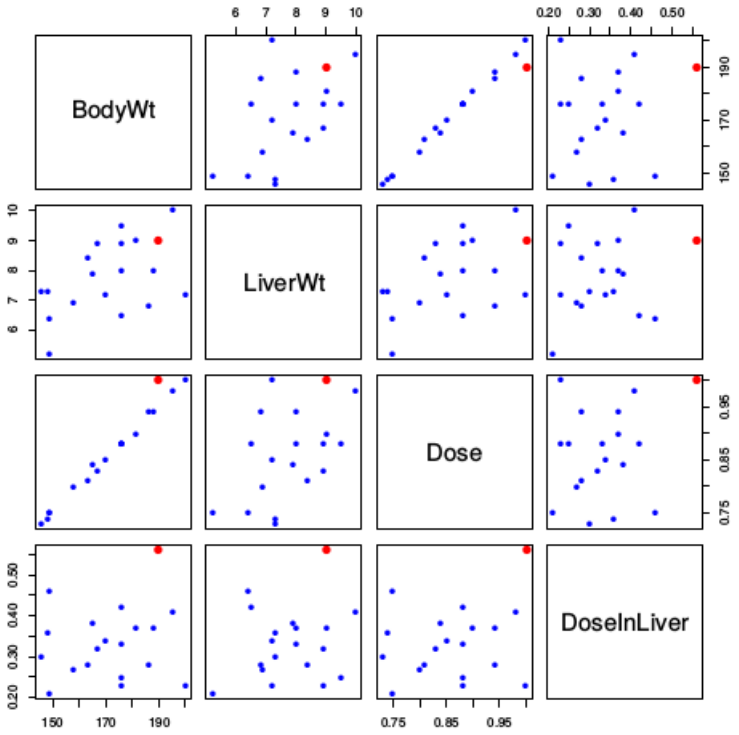
Volviendo al ejemplo de las ratas, vamos a ver porqué esa paradoja de los β_i distintos en regresión simple que en compuesta.

Calculando los potenciales y las distancias de Cook, vemos que:





La rata 3 es como muy rara, ya que tiene un potencial y una distancia de Cook muy altos. Vamos a pintarla en rojo, para que veamos de qué dato estamos hablando:



Vemos que algo raro sí parece, pero a simple vista no parece que podamos juzgar. Esta es la utilidad de estas medidas. Vamos a eliminar esa rata y recalcular los p-valores:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.311427	0.205094	1.518	0.151
BodyWt[-3]	-0.007783	0.018717	-0.416	0.684
LiverWt[-3]	0.008989	0.018659	0.482	0.637
Dose[-3]	1.484877	3.713064	0.400	0.695

Conclusión Una única observación hace que cambien totalmente las conclusiones respecto a si las variables son o no significativas.

2.3. Variable regresora cualitativa

Modelo unifactorial

Este es un tema que da para libros enteros. También se llama "diseño de experimentos". Nosotros vamos a ver el **Modelo unifactorial**, un caso concreto del tema para entender el fundamento matemático que hay detrás. Todo lo demás en lo que se profundiza en los libros son casos concretos.

Vamos a verlo con el siguiente ejemplo:

En un estudio para comparar la eficacia de tres fertilizantes se utiliza cada uno de ellos en 10 parcelas (asignando aleatoriamente cada parcela a uno de los tres fertilizantes) y posteriormente se registra el peso en toneladas de la cosecha resultante en cada parcela. Los datos son:

Fert. 1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
Fert. 2	3.07	3.29	4.04	4.19	3.41	3.75	4.87	3.94	6.28	3.15
Fert. 3	4.04	3.79	4.56	4.55	4.55	4.53	3.53	3.71	7.00	4.61

Una variable explicativa cualitativa se llama factor. Los valores que toma se llaman niveles. En este modelo los niveles son los distintos tratamientos que aplicamos a las unidades experimentales. En el ejemplo tenemos un factor (el tipo de fertilizante) que se presenta en tres niveles o tratamientos, que se aplican a las unidades experimentales (las parcelas).

Notación

- **k** es el número de niveles del factor (en este caso 3). En las transparencias se le denomina I.
- Y_{ij}, donde i es el nivel y j es la unidad dentro del nivel.
- $\overline{Y_{1\cdot}}$ es la media del nivel 1 (recordamos que $\overline{Y}_{\cdot j}$ sería la media de la columna).
- De esta manera, $\overline{Y_{\cdot\cdot}}$ es la media global de todos los datos.

Con esto, podemos entender los datos de la tabla:

Muestra	Respuestas				Medias	Desv. típicas
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$\bar{Y}_{1.}$	S_1
1	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$\bar{Y}_{2.}$	S_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
I	Y_{I1}	Y_{I2}	$\cdots$	Y_{In_I}	$\bar{Y}_{I.}$	S_I

2.3.1. Modelo unifactorial

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

donde

- μ_i es el nivel medio de la respuesta para el nivel i del factor.
- ε_{ij} es la variable de error que recoge el resto de variables que influyen en la respuesta.

  $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \rightarrow Y_{ij} \equiv N(\mu_i, \sigma^2)$.

  Vemos que son **homocedásticas**

Modelo unifactorial como caso de la regresión múltiple El modelo unifactorial se puede expresar en la forma $Y = X\beta + \varepsilon$, donde

- $Y = (Y_{1,1}, Y_{1,2}, Y_{k,n_k})'$
- $\beta = (\mu_1, ..., \mu_k)$
- $\varepsilon = (\varepsilon_{1,1}, \varepsilon_{1,2}, \varepsilon_{k,n_k})'$

Vamos a construir la matriz X:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \hline Y_{21} \\ \vdots \\ Y_{2n_2} \\ \hline \vdots \\ \hline Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix} = \left(\begin{array}{cccc} 1 & 0 & 0 & \dots \\ 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \\ 1 & 0 & 0 & \dots \\ \hline 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \\ 0 & 1 & 0 & \dots \\ \hline & & \vdots & \\ \hline 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 1 \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{array}\right) \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \hline \varepsilon_{k1} \\ \vdots \vdots \\ \varepsilon_{kn_k} \end{pmatrix}$$

Vamos a ver que esta matriz de diseño estima β correctamente:

$$\hat{\beta} = (\hat{\mu_1}, ..., \hat{\mu_n}) = \underbrace{(X'X)^{-1}}_{A} \underbrace{X'Y}_{B} = \underbrace{\begin{pmatrix} 1/n_1 & 0 & \dots & 0 \\ 0 & 1/n_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & 1/n_k \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} Y_{1\cdot} \\ Y_{2\cdot} \\ \vdots \\ Y_{k\cdot} \end{pmatrix}}_{B} = \begin{pmatrix} \overline{Y}_{1\cdot} \\ \overline{Y}_{2\cdot} \\ \vdots \\ \overline{Y}_{k\cdot} \end{pmatrix} = \overline{Y_{\cdot\cdot}}$$

Hipótesis ¿Cuál es el contraste equivalente que queremos realizar?

Queremos contrastar si el fertilizante es influyente, es decir: $H_0 : \mu_1 = ... = \mu_k$.

Modelos reducido y completo

Reducido

$$SCE_0 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = SCT$$

Completo

$$SCE = \sum_i \sum_j (Y_{ij} - \overline{Y}_{\cdot\cdot})^2 \overset{(1)}{=} \sum_i (n_i - 1) S_i^2$$

(1) ¿?¿?¿?

Teniendo esto, recordamos:

$$SCE_0 - SCE = SCT - SCE = SCR = \sum_i \sum_j (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = \sum n_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

Ahora ya podemos construir la tabla ANOVA:

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada	$\sum_{i=1}^k n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$I-1$	$\frac{\sum_{i=1}^k n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{k-1}$	F
Residual	$\sum_{i=1}^k (n_i - 1)S_i^2$	$n-k$	$\frac{\sum_{i=1}^k (n_i-1)S_i^2}{n-k}$	

¿Porqué $n-k$ grados de libertad? Si recordamos, ahí tendríamos $n-k-1 = n-(k+1) = n - \dim(V)$, siendo V el subespacio generado por la matriz X. En regresión múltiple teníamos $\dim(V) = k+1$, las k columnas más el término independiente. En este caso, nuestra matriz X tiene rango k.

Ejemplo: *Vamos a ver esto con los datos del ejemplo.*

Muestra	n_i	$\overline{Y_{i\cdot}}$	$S_{i\cdot}$
1	*10*	*5.445*	*0.976*
2	*10*	*3.999*	*0.972*
3	*10*	*4.487*	*0.975*

$$SCT = 10(5.445 - 4.644)^2 + 10(3.999 - 4.644)^2 + 10(4.487 - 4.644)^2 \simeq 10.82$$

$$SCE = 9(0.976)^2 + 9(0.972)^2 + 9(0.975)^2 = 25.62$$

Ahora ya podemos construir el estadístico:

$$\frac{SCR/(k-1)}{SCE/(n-k)} = \ldots = 5.702$$

Por otro lado, $F_{2,27;0.05} = 3.35$

Como el valor obtenido es mayor que el estadístico, rechazamos la hipótesis.

```
> cosecha = read.table('cosecha.txt',header=TRUE)
> resultado = aov(cosecha ~ factor(fertilizante), data=cosecha)
> summary(resultado)

                    Df SumSq MeanSq F value Pr(>F)
factor(fertilizante) 2 10.82  5.411   5.702 0.00859 **
Residuals           27 25.62  0.949
```

Capítulo III

Clasificación

Disponemos de una muestra de k variables medidas en n unidades u objetos que pertenecen a dos grupos o poblaciones *(training data)*.

Cada observación $i = 1, \ldots, n$ consiste en un vector $(x_i', y_i)'$, donde $x_i \in \mathbb{R}^k$ son las k variables e $y \in \{0, 1\}$ indica el grupo al que pertenece la unidad en la que se han obtenido.

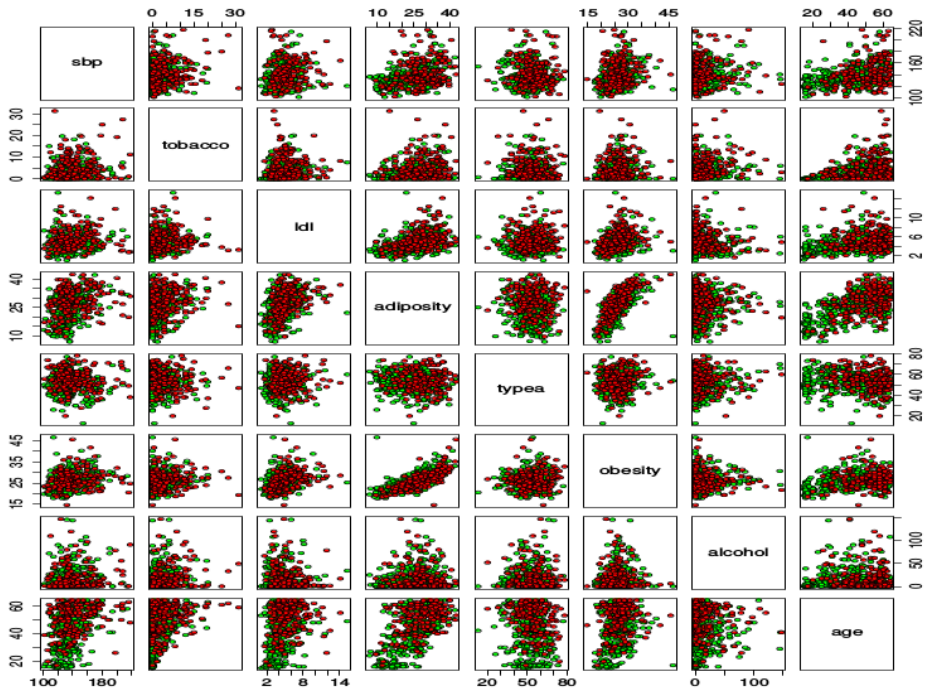
Objetivo: Asignar una nueva unidad con valores x (e y desconocida) a uno de los dos grupos (**obtener una regla de clasificación**).

Este problema tiene diferentes nombres en la literatura en inglés: "supervised classification", "statistical learning", "discrimination", "machine learning", "pattern recognition", etc.

Vamos a ver un ejemplo, para entender mejor el tema

Ejemplo: *En un estudio de factores de riesgo en enfermedades coronarias, se dispone de datos de 462 personas (de las que 160 habían sufrido infartos y 302 eran controles). Para cada una de ellas se midieron las siguientes variables:*

Nombre variable	*Descripción*
`sbp`	*Tensión sanguínea sistólica*
`tobacco`	*Consumo de tabaco*
`ldl`	*Colesterol*
`adiposity`	*Medida de adiposidad*
`typea`	*Comportamiento "tipo A"*
`obesity`	*Medida de la obesidad*
`alcohol`	*Consumo de alcohol*
`age`	*Edad*



Ahora nos gustaría poder predecir si un nuevo individuo va a tener un infarto o no en función de su consumo de tabaco, colesterol, ...

1. Regla de Mahalanobis

La idea es asignar el nuevo individuo al grupo cuyo centro es más cercano (cercano en el sentido de la distancia de Mahalanobis).

Regla de Mahalanobis

Definición 1.1 **Regla de Mahalanobis**. *Para $i = 0, 1$ denotamos P_i a la distribución condicionada $X|Y = i$. Suponemos que P_i es una distribución con vector de medias μ_i y matriz de covarianzas Σ_i.*

Regla de Mahalanobis: *Clasificar x en el grupo 1 (i.e. $Y = 1$) si y solo si*

$$(x - \mu_0)'\Sigma_0^{-1}(x - \mu_0) > (x - \mu_1)'\Sigma_1^{-1}(x - \mu_1).$$

En la práctica se usan los vectores de medias y las matrices de covarianzas muestrales, ya que no disponemos de los reales.

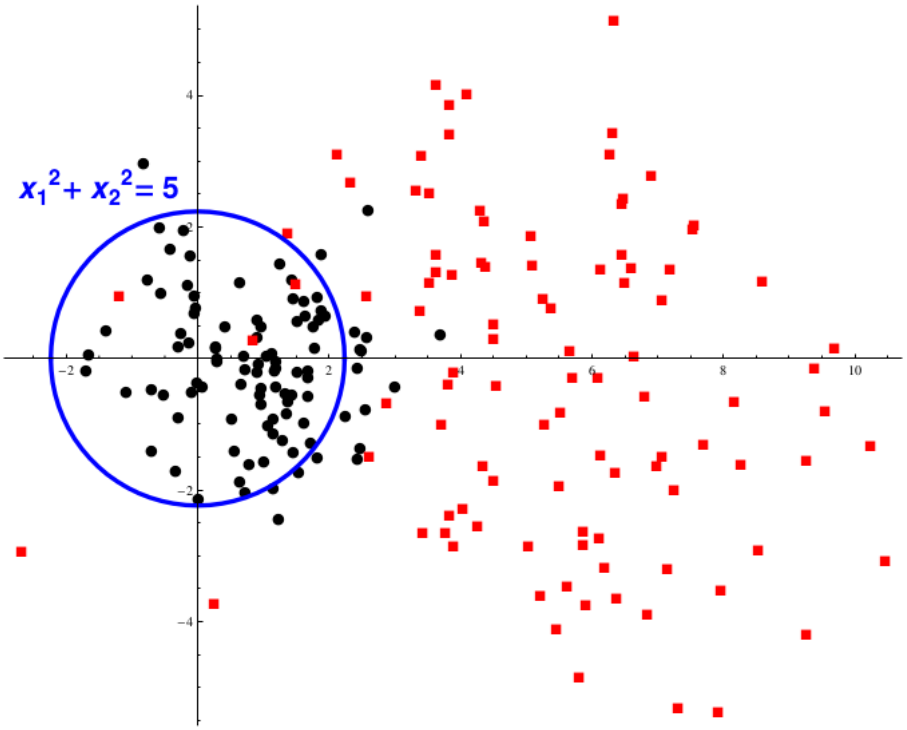
La frontera de clasificación sería cuando

$$(x - \mu_0)'\Sigma_0^{-1}(x - \mu_0) = (x - \mu_1)'\Sigma_1^{-1}(x - \mu_1).$$

Esta frontera será una curva cuadrática (por ser una igualdad entre formas cuadráticas).

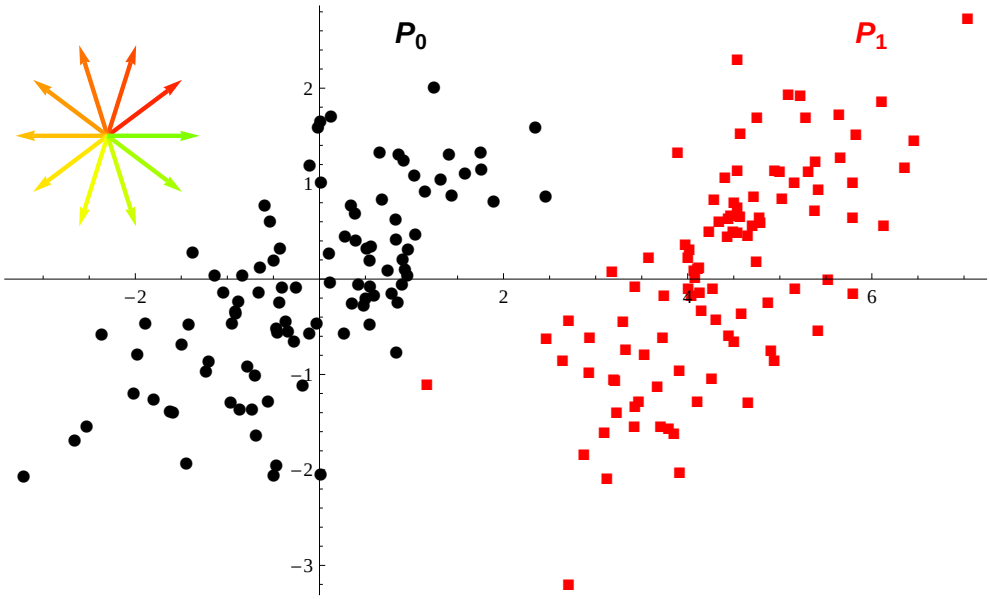
$$\mu_0 = (1, 0)', \mu_1 = (5, 0)', \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}.$$

Ejemplo:

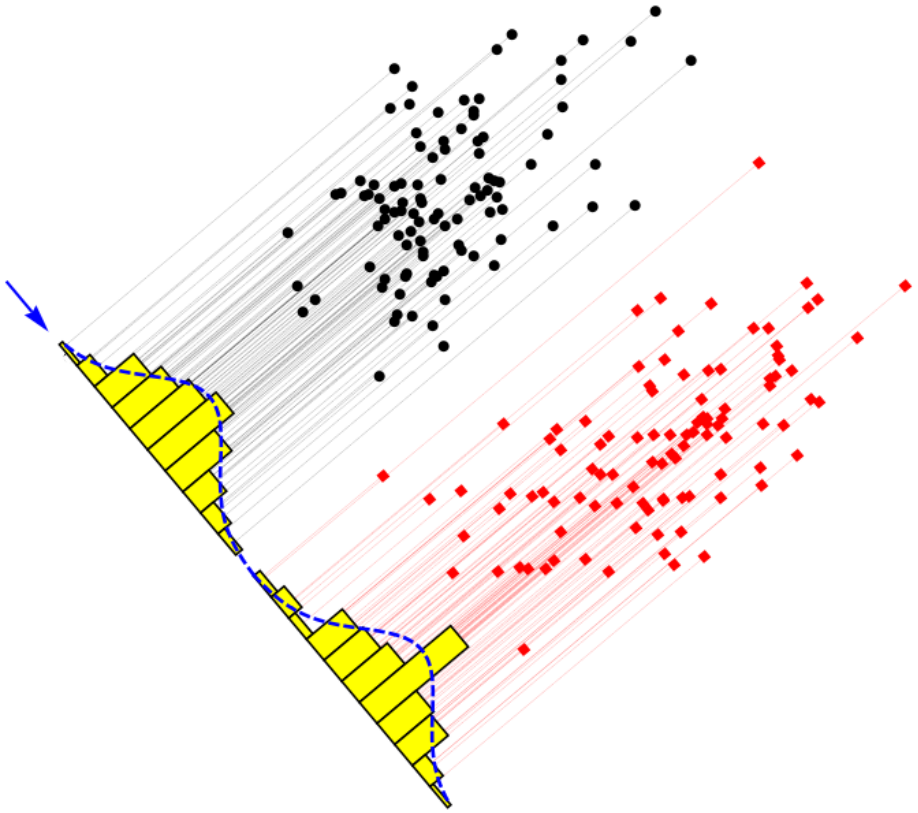


2. Regla de Fisher

Vamos a suponer $\Sigma_1 = \Sigma_0$, que es cuando mejor funciona esta regla. Vamos a suponer que tenemos estos datos:

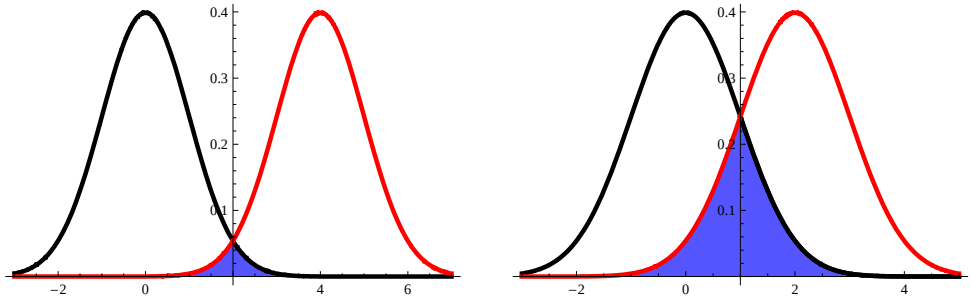


Y la idea intuitiva es construir una recta sobre la que proyectar, para construir 2 histogramas. Si nos dan un nuevo dato, lo clasificaremos en el histograma al que sea más probable que pertenezca.

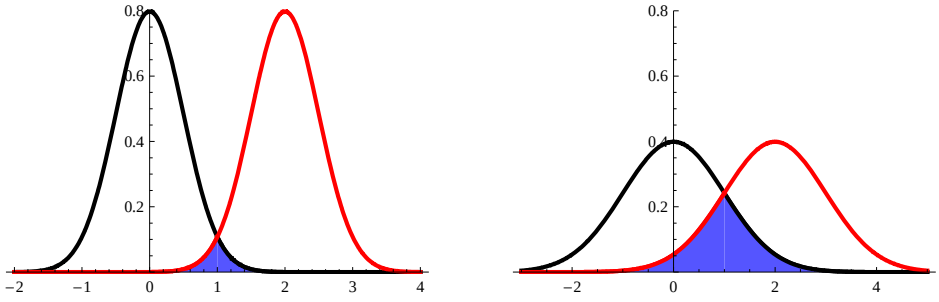


¿Cómo construir esta recta? Tenemos que tener en cuenta 2 cosas:

- Una buena dirección debe separar bien los centros de los grupos. La distancia entre las medias $(a'\mu_0 - a'\mu_1)^2 = a'Ba$, donde $B = (\mu_0 - \mu_1)(\mu_0 - \mu_1)'$, debe ser grande.



- La varianza de las proyecciones dentro de los grupos ($a'\Sigma a$) debe ser lo menor posible.



Cociente de Rayleigh

La idea de Fisher fue maximizar el ratio de la separación de los centros entre las varianzas, es decir, maximizar el **Cociente de Rayleigh**

$$f(a) = \frac{a'Ba}{a'\Sigma a}$$

para cualquier dirección $a \in \mathbb{R}^n$.

Observación: Este problema tiene infinitas soluciones, ya que $f(a) = f(\lambda a) \forall \lambda \in \mathbb{R}$. Para solucionar esto, imponemos la normalización tal que $a'\Sigma a = 1$ (aunque no sea la única).

Vamos a calcular el máximo de esta función[1]

$$\nabla f(a) = \frac{2Ba(a'\Sigma a) - 2\Sigma a(a'Ba)}{(a'\Sigma a)^2} \to \nabla f(\omega) = 0 \iff B\omega(\omega'\Sigma\omega) = \Sigma\omega(\omega'B\omega)$$

Si llegamos a encontrar ese ω, tendrá estas 2 propiedades

- $B\omega = (\mu_0 - \mu_1)\underbrace{(\mu_0 - \mu_1)'\omega}_{\alpha}$ siendo α un escalar. Esto quiere decir que $B\omega = (\mu_0 - \mu_1)\alpha$, esto es: $B\omega$ es proporcional a $\mu_0 - \mu_1$.
- $\Sigma\omega$ es proporcional a $\Sigma^{-1}(\mu_0 - \mu_1)$

Regla de Fisher

Definición 2.1 **Regla de Fisher**. *Clasificar x en el grupo 1 (i.e. $Y = 1$) si y solo si*

$$\omega'\left(x - \frac{\mu_0 + \mu_1}{2}\right) > 0$$

donde $\omega = \Sigma^{-1}(\mu_1 - \mu_0)$.

En caso de tener $\omega'\left(x - \frac{\mu_0+\mu_1}{2}\right) < 0$, clasificaríamos en el grupo 2.

Observación: Si en la regla de Mahalanobis se supone $\Sigma_1 = \Sigma_2$, se obtiene la regla de Fisher:

$$\begin{aligned}
\omega'\left(x - \frac{\mu_0 + \mu_1}{2}\right) &> 0 \\
(\mu_1 - \mu_0)'\Sigma^{-1}\left(x - \frac{\mu_0 + \mu_1}{2}\right) &> 0 \\
(\mu_1 - \mu_0)'\Sigma^{-1}x - (\mu_1 - \mu_0)'\Sigma^{-1}\left(\frac{\mu_0 + \mu_1}{2}\right) &> 0 \\
(\mu_1 - \mu_0)'\Sigma^{-1}x - \underbrace{\mu_1'\Sigma^{-1}\frac{\mu_0}{2}}_{A} + \mu_0'\Sigma^{-1}\frac{\mu_0}{2} - \mu_1'\Sigma^{-1}\frac{\mu_1}{2} + \underbrace{\mu_0'\Sigma^{-1}\frac{\mu_1}{2}}_{B} &> 0
\end{aligned}$$

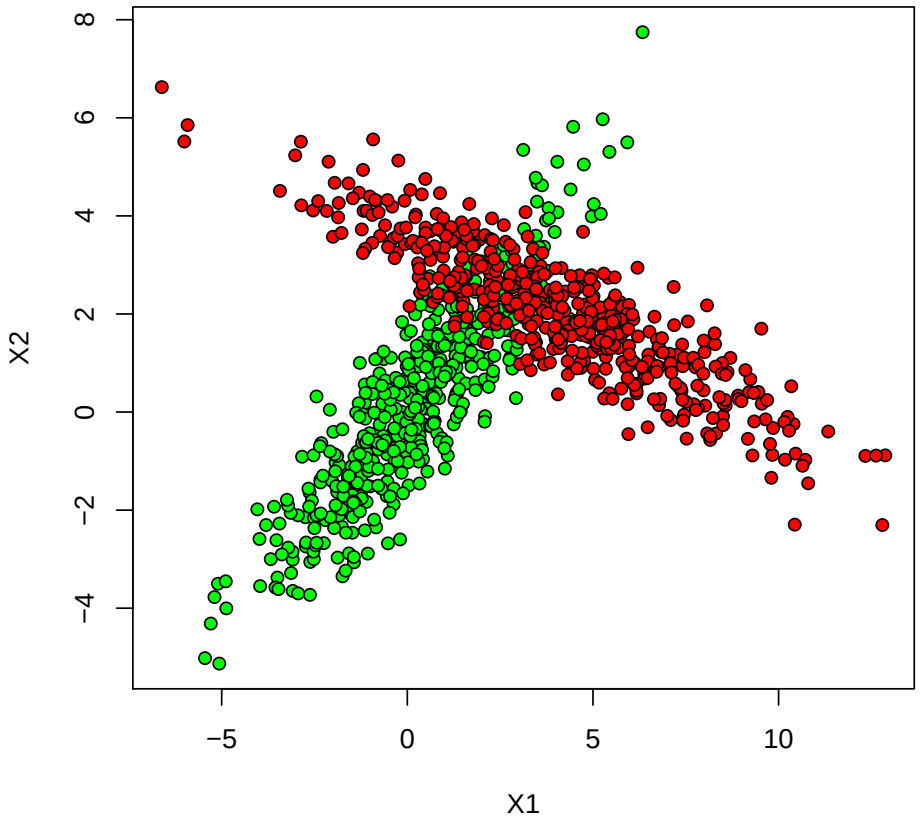
[1] Utilizamos $\nabla a'\Sigma a = 2a\Sigma$. La derivada de una forma cuadrática que es algo que ya deberíamos haber sabido de otras asignaturas del grado.

Sabiendo $A, B \in \mathbb{R} \implies A = A^T = B$ podemos seguir:

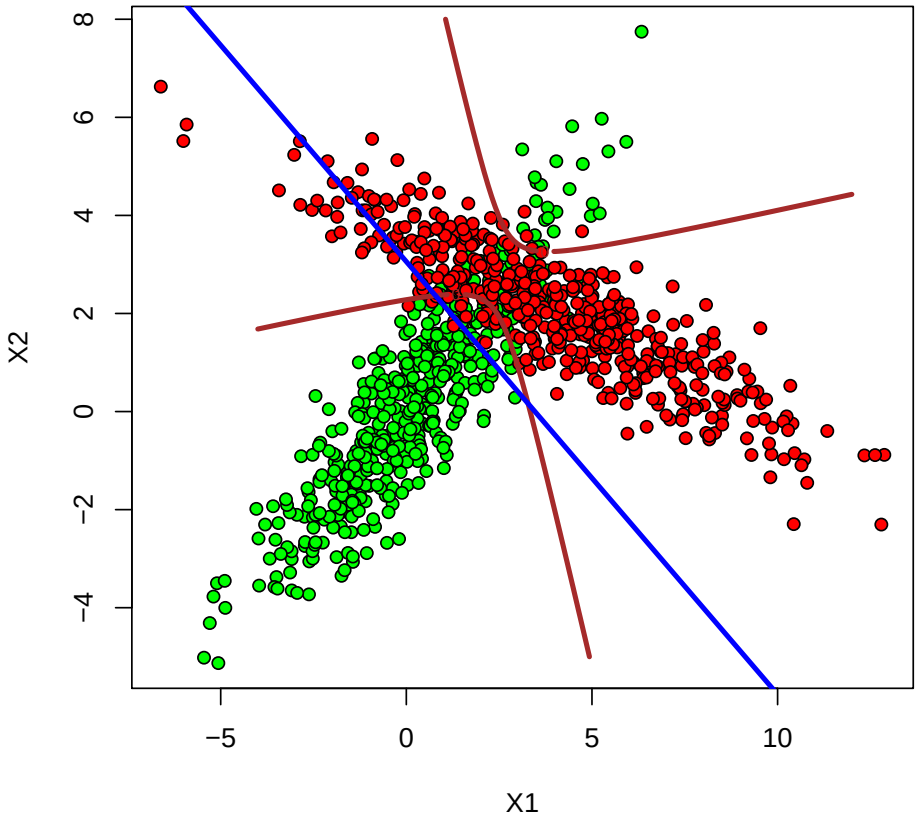
$$\begin{aligned}
2(\mu_1 - \mu_0)'\Sigma^{-1}x + \mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1 > 0 \\
x'\Sigma^{-1}(\mu_1 - \mu_0) + (\mu_1 - \mu_0)'\Sigma^{-1}x + \mu_0'\Sigma^{-1}\mu_0 - \mu_1'\Sigma^{-1}\mu_1 > 0 \\
x'\Sigma^{-1}x - x'\Sigma^{-1}\mu_0 - \mu_0'\Sigma^{-1}x + \mu_0'\Sigma^{-1}\mu_0 > x'\Sigma^{-1}x - x'\Sigma^{-1}\mu_1 - \mu_1'\Sigma^{-1}x + \mu_1'\Sigma^{-1}\mu_1 \\
(x'\Sigma^{-1} - \mu_0'\Sigma^{-1})(x - \mu_0) > (x'\Sigma^{-1} - \mu_1'\Sigma^{-1})(x - \mu_1) \\
(x - \mu_0)'\Sigma^{-1}(x - \mu_0) > (x - \mu_1)'\Sigma^{-1}(x - \mu_1)
\end{aligned}$$

Por tanto hemos llegado a la regla de Mahalanobis desde la regla lineal de Fisher suponiendo que $\Sigma_1 = \Sigma_2$.

Observación: Esta regla funciona bien cuando $\Sigma_0 = \Sigma_1$, es decir, las nubes de puntos tienen la misma forma y orientación. ¿Qué ocurre cuando las matrices de covarianzas son distintas?

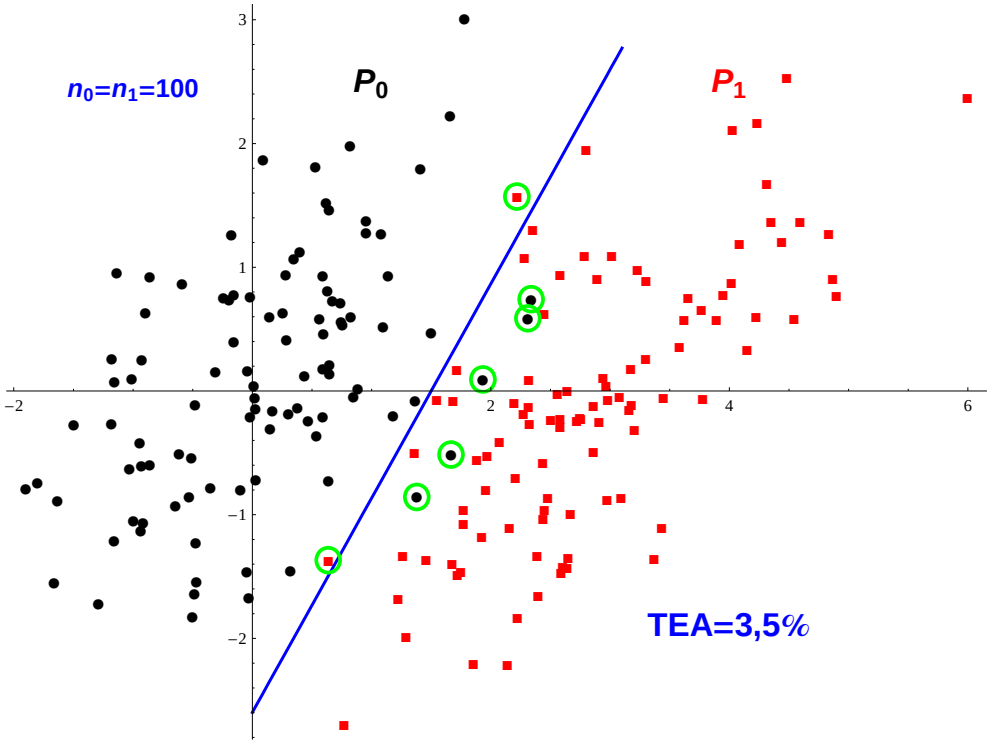


En la imagen vemos que la regla de Fisher (recta azul) no divide nada bien los datos. Por otro lado, vemos la división por la distancia de Mahalanobis (en rojo) que es mucho más adecuada.



2.1. Validación del modelo

Acabamos de ver un ejemplo de un caso en el que vemos (a simple vista) que un modelo clasifica mejor que el otro. ¿Porqué sabemos que es mejor? Porque comete menos errores. Vamos a ver en esta sección cómo calcular errores.



Tasa de error aparente

Definición 2.2 **Tasa de error aparente**.

$$TEA := \frac{\textit{Total de mal clasificados en la muestra}}{n} 100\,\%.$$

El problema de esta tasa de error es que infraestima la tasa de error. Esto se debe a que estamos utilizando los mismos puntos para construir el modelo y para evaluarlo.

La solución es la creación de particiones de train y de test. De esta manera utilizamos los datos de train para construir el modelo y el de test para evaluarlo. ¿Y cómo asegurar que la tasa de error no depende de las particiones construidas? No se puede, pero lo que podemos hacer es utilizar la Validación cruzada.

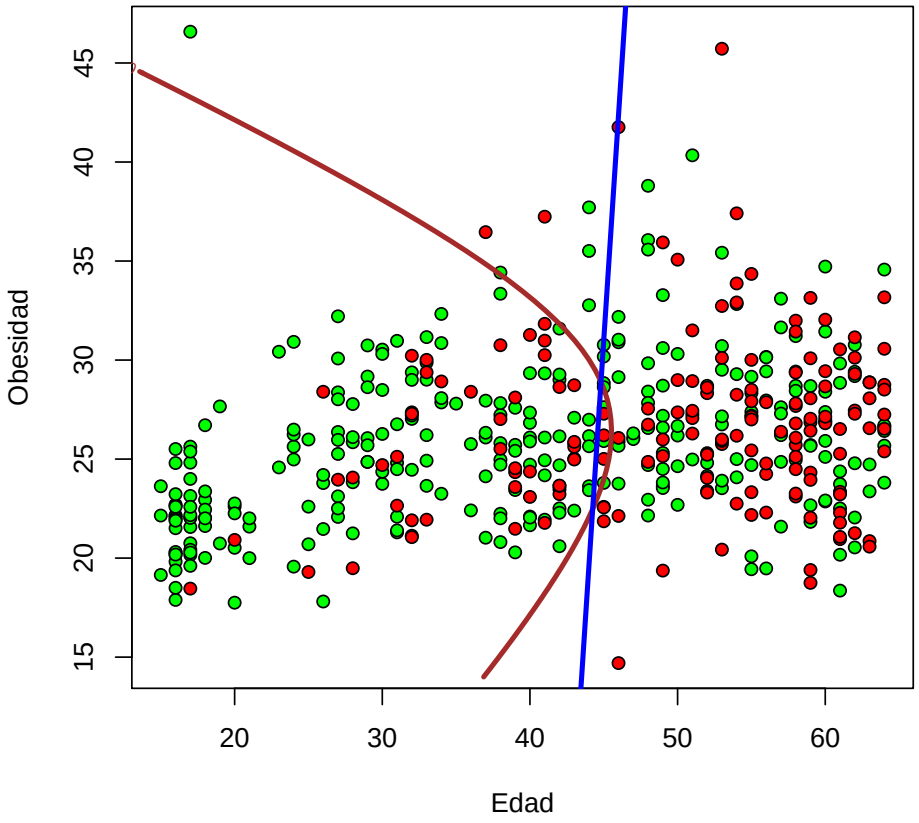
Validación cruzada

Definición 2.3 **Validación cruzada**.*Omitimos un dato de los n observados y generamos la regla de clasificación con los $n-1$ restantes. Clasificamos la observación apartada y repetimos el procedimiento para cada una de las observaciones.*

3. Regresión logística

Vamos a recordar el ejemplo de infartos de miocardio y vamos a ver cómo clasifican los modelos de clasificación estudiados hasta ahora.

En rojo vemos utilizando la distancia de Mahalanobis y en azul la regla de Fisher.



Vemos que ninguno de los 2 es bueno. Por ello, vamos a ver otro método de clasificación planteando la clasificación como un problema de regresión múltiple.

No podemos utilizarlo como tal porque la variable regresora Y_i es binaria y en la regresión múltiple era continua (más concretamente normal). Por ello, hay que buscar una alternativa.

3.1. Construcción del modelo

Las variables $Y_1, \dots, Y_n$ son independientes y tienen distribución de Bernoulli.

Denotamos $p_i = \mathbb{P}(Y_i = 1 \mid x_i)$. La probabilidad de "éxito" depende de las variables regresoras.

Una relación lineal $p_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ no es adecuada (¿por qué?)

Suponemos que la relación entre p_i y x_i viene dada por

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}}},$$

es decir,

$$p_i = F(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}),$$

Función logística

donde $F(x) = 1/(1 + e^{-x})$ es la **Función logística**.

Observación:

- $$F(-x) = \frac{1}{1+e^{x}} = \frac{e^{-x}}{1+e^{-x}} = 1 - F(x)$$
- $$F'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = F(x)(1-F(x))$$
- $$\log \frac{p_i}{1-p_i} = \beta' x_i$$

Razón de probabilidades

Definición 3.1 **Razón de probabilidades**. *Llamamos O_i a la **razón de probabilidades** para la observación i:*

$$O_i = \frac{p_i}{1-p_i}$$

Interpretación El estado de las apuestas.

¿Cómo varía la razón de probabilidades si la variable regresora x_{ij} se incrementa una unidad? Si se cumple el modelo de regresión logística, entonces

$$O_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}}$$

Con lo que la variación:

$$\frac{O_i'}{O_i} = \frac{e^{\beta_0 + \cdots + \beta_j (x+1) + \cdots + \beta_k x_{ik}}}{e^{\beta_0 + \cdots + \beta_j x + \cdots + \beta_k x_{ik}}} = e^{\beta_j}.$$

Por tanto e^{β_j} es la variación de la razón de probabilidades cuando la variable regresora j se incrementa en una unidad y el resto de variables permanece constante.

3.2. Estimación de los parámetros

Para estimar los parámetros se usa el método de máxima verosimilitud.

Por ejemplo, si observamos los datos

x_i	2	1	3
Y_i	0	1	1

entonces $\hat{\beta}_0$ y $\hat{\beta}_1$ son los valores que maximizan la función de verosimilitud

$$L(\beta_0, \beta_1) = P(Y=0 \mid x=2) P(Y=1 \mid x=1) P(Y=1 \mid x=3)$$
$$L(\beta_0, \beta_1) = \left(1 - \frac{1}{1+e^{-\beta_0 - 2\beta_1}}\right) \left(\frac{1}{1+e^{-\beta_0 - \beta_1}}\right) \left(\frac{1}{1+e^{-\beta_0 - 3\beta_1}}\right)$$

Se suelen aplicar métodos numéricos estándar de optimización ya que es difícil sacar los valores exactos.

La fórmula general es:

$$L(\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

De esta manera, cuando $Y_i = 0$, entonces nos quedamos con el término $(1-p_i)$. Por otro lado, cuando $Y_i = 1$, tenemos el término p_i.

Vamos a derivar el logaritmos de la función de verosimilitud:

$$l(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} \left[Y_i \log p_i + (1-Y_i)\log(1-p_i) \right]$$

$$\nabla l(\beta) = \sum_{i=1}^{n} \left[\frac{Y_i}{p_i} p_i(1-p_i)\mathbf{x}_i - \frac{1-Y_i}{1-p_i} p_i(1-p_i)\mathbf{x}_i \right] = \ldots = \sum_{i=1}^{n} (y_i - p_i)\mathbf{x}_i$$

En esta última cuenta podemos ver la conveniencia de la función logística. En realidad, podríamos utilizar cualquier otra función $F : \mathbb{R} \to [0,1]$. Utilizando otras funciones, no se tendrían estas simplificaciones que hacen relativamente fácil el cálculo del E.M.V. de β

Conclusión: $\hat{\beta}$ verifica:

$$\sum (Y_i - \hat{p}_i)\mathbf{x}_i = 0$$

donde:

$$\hat{p}_i = \frac{1}{1 + e^{-\hat{\beta}'\mathbf{x}_i}}$$

Relación con regresión simple En regresión lineal simple, teníamos que el E.M.V. verifica:

$$\sum (Y_i - \hat{Y}_i)x_i = 0$$

donde $\hat{Y}_i = \hat{\beta}' x_i$.

Observación:

1. Recordamos que en el caso unidimensional, el E.M.V. tiene una distribución de media el valor verdadero y varianza $\frac{1}{I_F}$ donde I_F es la información de Fisher y se calculaba derivando 2 veces la verosimilitud (más o menos)

   En el caso multidimensional, tenemos:

$$\hat{B} \equiv N_{k+1}\left(\beta, (X'\hat{W}X)^{-1}\right)$$

Teniendo

$$\hat{W} = \begin{pmatrix} \hat{p_1}(1-\hat{p_1}) & 0 & \dots & \dots \\ 0 & \hat{p_2}(1-\hat{p_2}) & \dots & \dots \\ \vdots & \ddots & & \vdots \\ 0 & \dots & 0 & \hat{p_n}(1-\hat{p_n}) \end{pmatrix}$$

y siendo X la matriz de diseño.

2. Desviaciones:

$$D_i^2 = -2\left[Y_i \log \hat{p}_i + (1-Y_i)\log(1-\hat{p}_i)\right]$$

D_i mide la bondad de ajuste del modelo a la observación i.

$Y_i = 1 \to D_i^2 = -2\log \hat{p}_i$. De esta manera, si $\hat{p}_i \to 0$, entonces $D_i^2 \to \infty$, cosa que tiene toda la lógica del mundo. Si el modelo me da muy poca probabilidad de clasificarlo como enfermo, siendo enfermo entonces estoy muy desviado y el modelo se ajusta mal para esa observación. Si por el contrario, $\hat{p}_i \to 1 \implies D_i^2 \to 0$, es decir, si me da mucha probabilidad, entonces hay muy poca desviación.

3. Todas las consideraciones anteriores sobre las desviaciones se podrían haber hecho sin multiplicarlo por el 2. Vamos a ver porqué aparece.

   Maximizar $l(\beta)$ es lo mismo que minimizar $-2 \cdot l(\beta)$.

Desviación residual

Definición 3.2 **Desviación residual**. *Vemos que cada sumando de $l(\beta)$ el D_i^2 correspondiente multiplicado por -2, es decir:*

$$-2l(\beta) = \sum D_i^2 = D^2$$

Este sumatorio es la desviación residual

Observación: Igual que en regresión múltiple minimizábamos la suma de los residuos, en este, el E.M.V. minimiza la desviación residual. Análogamente, si el modelo tiene demasiadas variables esta desviación se reduce y provoca que demasiadas variables reduzcan este valor. Lo mismo ocurría con el coeficiente de determinación ajustado.

La solución es construir una "desviación residual ajustada" que penalice los modelos complicados.

Criterio de Información de Akaike (AIC)

Utilizamos el **Criterio de Información de Akaike (AIC)**

$$AIC = D^2 + 2(k+1)$$

Demostración. Si alguien está interesado en saber cuál es la justificación teórica del sumando "2(k+1)", puede buscar el artículo de Akaike en el que se explica, porque es algo que ni el profesor sabe bien. □

```
# Numero de observaciones
n = 100

# Parametros
beta0 = 0
beta1 = 3

# Genera los datos
x = rnorm(n)
p = 1/(1+exp(-beta0-beta1*x))
y = rbinom(n,1,p)

# Ajusta el modelo
reg = glm(y~x,family=binomial)
summary(reg)

# # # # Y obtenemos: # # # #


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8340  -0.7598  -0.1568   0.7623   2.4224

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.1157     0.2585  -0.447    0.655
x             2.7083     0.5854   4.627 3.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 137.628  on 99  degrees of freedom
Residual deviance:  90.396  on 98  degrees of freedom
AIC: 94.396

Number of Fisher Scoring iterations: 5
```

- *Los errores típicos son los elementos de la diagonal de la matriz $X'\hat{W}X$.*
- *Ejemplos:*

  $H_0 : \beta_0 = 0$

  Vemos el valor -0.447 y lo comparamos con la tabla de la normal, o en su defecto cogemos el p-valor 0.655 y no rechazamos la hipótesis (cosa que está bien)

  Para el $IC(\beta_1)$

$$IC(\beta_1) = [2.70 \mp \underbrace{z_{\alpha/2}}_{1.96} 0.5859] =$$

  ¿Y porqué utilizamos la N y no una t como siempre?
- *Como en este caso $k = 1$, vemos que $AIC = 4 + (\text{Resudial deviance}) = 4 + D^2$*
- *¿Qué es la "Null deviance"?*

Null deviance

Definición 3.3 ***Null deviance****. Tomando $H_0 : \beta_1 = ... = \beta_k = 0$, la "Null deviance" es D_0^2 para este modelo reducido.*

- *Family = binomial es el valor que utilizar para que sea regresión logística.*

3.3. Contraste de un modelo reducido

Modelo completo (M) y un modelo reducido (M_0) que procede de imponer p restricciones en el modelo completo y tenemos el contraste $H_0 : M_0$ es cierto.

Vamos a estudiar $D_0^2 - D^2$ (que siempre es negativo, ya que al aumentar el número de variables el máximo del E.M.V. es mayor)

$$D_0^2 - D^2 = -2\log L(\hat{\beta}^{(0)}) + 2\log L(\hat{\beta}) = -2\log \frac{L(\hat{\beta}^{(0)})}{L(\hat{\beta})}$$

Cociente de verosimilitudes

Esto es un **Cociente de verosimilitudes**.

Llamamos $\Lambda = \frac{L(\hat{\beta}^{(0)})}{L(\hat{\beta})}$ y además sabemos que

$$-2\log\Lambda \xrightarrow[H_0]{d} \chi_p^2$$

¿Porqué sabemos que se distribuye como una χ^2? Este es un resultado visto en [Julián Moreno, 2013, IV.4.3].

Entonces, la región de rechazo es:

$$R = \{D_0^2 - D^2 > \chi^2_{p;\alpha}\}$$

Vamos a verlo en el ejemplo de la salida de R.

$D_0 = 137.628$, con lo que $D_0^2 - D^2 = 47.2$ y $p = 99 - 98 = 1$, entonces:

$$R = \{47.2 > \chi^2_{1;0.05}\} \rightarrow \text{ p-valor } \simeq 0$$

Test de Wald
Test de razón de verosimilitudes

Conclusión Para contrastar $H_0 : \beta_j = 0$ tenemos 2 posibilidades. La primera es el **test de Wald** que es el que hace R pero también podemos construir un **test de razón de verosimilitudes**, que es lo que acabamos de hacer. Como en el ejemplo de R hay 2 coeficientes, M_0 es sólamente $\beta_1 = 0$.

Ejemplo: *Para el siguiente resultado:*

```
reg1 = glm(Y ~ edad, family='binomial')
summary(reg1)


Call:
glm(formula = Y ~ edad, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4321  -0.9215  -0.5392   1.0952   2.2433
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.521710   0.416031  -8.465  < 2e-16 ***
edad         0.064108   0.008532   7.513 5.76e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 525.56  on 460  degrees of freedom
AIC: 529.56

Number of Fisher Scoring iterations: 4
```

Queremos calcular:

- *Modelo ajustado* $P(y = \hat{1}|edad)$
- *Interpretación.*
- $H_0 : \beta_1 = 1$ *con los 2 métodos conocidos, Wald y Razón de verosimilitudes (RV).*
- *IC para* e^{β_1}

Vamos a verlo.

Recordamos que el modelo es:

$$P_i = P(Y_i = 1|X = x) = \frac{1}{1 + e^{-\beta' x}}$$

donde $Y_1, ..., Y_n \sim B(1, p_i)$ *independientes.*

Entonces,

$$P(y = 1|\hat{EDAD}) = \frac{1}{1 + \exp(3.5217 - 0.0651 \cdot EDAD)}$$

La ***interpretación*** *tiene que ver con* $e^{\hat{\beta}_1} = 1.066$*. Si la edad se incrementa un año, ¿Cuánto aumentan las apuestas de tener un infarto?* $\tilde{O}_i = 1.066 \cdot O_i$

Para el ***contraste*** *con el método de* ***Wald****, rechazamos* H_0 *porque el p-valor del contrate* $\beta_1 = 0$ *es muy pequeño.*

Para el ***contraste*** *utilizando RV, tenemos:* $D_0^2 - D^2 = 591.11 - 525.56 = 70.55$ *y utilizamos el resultado que asintóticamente es una* χ^2*, con lo que la región de rechazo:*

$$R = \{D_0^2 - D^2 > \chi^2_{1;\alpha}\}$$

Consultando las tablas, vemos que para casi cualquier α*, rechazamos porque el p-valor (que no estamos calculando) debe de ser casi 0.*

Vamos a construir el ***intervalo de confianza****.*

$$IC_{0.95}(\beta_1) = [0.064 - 1.96 \cdot 0.0085]$$

Con lo que:

$$IC_{0.95}(e^{\beta_1}) = [e^{0.064 - 1.96 \cdot 0.0085}, e^{-0.064 + 1.96 \cdot 0.0085}]$$

Observación: Como hemos visto, tenemos 2 contrastes para una misma hipótesis. Son contrastes distintos que usualmente dan los mismo resultados. Se podrían "cocinar" un poco los datos para rechazar con un contraste y aceptar con el otro.

3.4. Con 2 variables regresoras

Ahora vamos a tomar el modelo con la edad y la obesidad. En este caso, el estadístico $D_0^2 - D^2 = 596.11 - 525.55$ sirve para contrastar $H_0 : \beta_1 = \beta_2 = 0$ y se compara con χ^2_2 (2 grados de libertad porque imponemos 2 restricciones).

Vamos a verlo:

```
1
2 reg = glm(Y ~ edad+obesidad, family='binomial')
3 # # Recordamos:
4 # reg1 = glm(Y ~ edad, family='binomial')
5
6 summary(reg)
7
8 Deviance Residuals:
9     Min       1Q   Median       3Q      Max
10 -1.4401  -0.9227  -0.5384   1.0905   2.2497
11
12 Coefficients:
13             Estimate Std. Error z value Pr(>|z|)
14 (Intercept) -3.581465   0.742611  -4.823 1.42e-06 ***
15 edad          0.063958   0.008674   7.374 1.66e-13 ***
16 obesidad      0.002523   0.025934   0.097    0.923
17 ---
18 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' 1
19
20 (Dispersion parameter for binomial family taken to be 1)
21
22     Null deviance: 596.11  on 461  degrees of freedom
23 Residual deviance: 525.55  on 459  degrees of freedom
24 AIC: 531.55
25
26 Number of Fisher Scoring iterations: 4
```

Ahora comparamos si al añadir la obesidad obtenemos suficiente información.

```
1 anova(reg1, reg, test='Chisq')
2 Analysis of Deviance Table
3
4 Model 1: Y ~ edad
5 Model 2: Y ~ edad + obesidad
6   Resid. Df Resid. Dev Df   Deviance P(>|Chi|)
7 1       460     525.56
8 2       459     525.55  1 0.0094552     0.9225
```

Al ser las desviaciones muy similares quiere decir que al añadir la obesidad no obtenemos suficiente información como para que merezca la pena añadirla. Es por ello que el p-valor es muy grande.

4. Regresión logística como clasificador

Regla de clasificación logística

Definición 4.1 **Regla de clasificación logística**. *Clasificar x en el grupo 1 (i.e. $Y = 1$) si y*

solo si

$$\widehat{\mathbb{P}(Y=1|x)} > \widehat{\mathbb{P}(Y=0|x)}$$

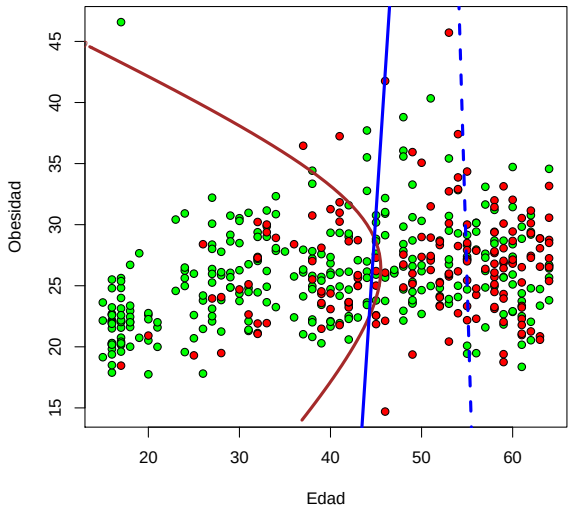
Sustituyendo por la función logística tenemos una regla lineal (diferente en general a la de Fisher): Clasificar x en el grupo 1 (i.e. $Y=1$) si y solo si

$$\hat{\beta}_0 + \hat{\beta}_k x_1 + \ldots + \hat{\beta}_k x_k > 0.$$

En el **ejemplo**, clasificamos a un individuo como enfermo si y solo si

$$-3.58 + 0.064 \cdot \texttt{edad} + 0.0025 \cdot \texttt{obesidad} > 0.$$

En la nube de puntos, incluimos la frontera de clasificación según la regla de regresión logística como línea discontinua azul:



5. Regla de Bayes

Vamos a ver (en la tercera asignatura del grado) la regla de Bayes para clasificar.

Imaginemos que tenemos las probabilidades reales, entonces, $P(Y=1|x) > P(Y=0|x)$.

En el caso en que

1. P_0 tiene densidad f_0 y P_1 tiene densidad f_1,
2. las **probabilidades a priori** de las poblaciones son

$$\mathbb{P}(P_0) = \pi_0, \quad \mathbb{P}(P_1) = \pi_1 \quad (\pi_0 + \pi_1 = 1).$$

3. Distribución condicionada de las variables explicativas: $P(X|Y=0) = 0$ tiene densidad f_0 y $P(X|Y=1)$ tiene densidad f_1.

   Ahora, aplicamos la regla de bayes:

$$P(Y=1|X=x) = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} = \frac{f_1(x)\pi_1}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

Por otro lado:

$$P(Y=0|X=x) = \frac{P(X=x|Y=0)P(Y=0)}{P(X=x)} = \frac{f_0(x)\pi_0}{\pi_0 f_0(x) + \pi_1 f_1(x)}$$

Conclusión

$$\mathbb{P}(Y=1|x) > \mathbb{P}(Y=0|x) \Leftrightarrow \pi_1 f_1(x) > \pi_0 f_0(x).$$

Proposición 5.1. *La regla Bayes es óptima (su error de clasificación es el mínimo posible).*

Demostración. No da tiempo a verlo, asique nos lo creemos. □

5.1. Bayes para normalidad

Supongamos que f_0 y f_1 son normales: para $x \in \mathbb{R}^p$,

$$f_i(x) = |\Sigma_i|^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}, \quad i=0,1.$$

Regla Bayes bajo normalidad
Se clasifica x en P_0 si:

$$d^2_{M_0}(x,\mu_0) < d^2_{M_1}(x,\mu_1) - 2\log\left(\frac{\pi_1 |\Sigma_1|^{-\frac{1}{2}}}{\pi_0 |\Sigma_0|^{-\frac{1}{2}}}\right)$$

Demostración. Sabemos que clasificamos x en $P_0 \Leftrightarrow \pi_0 f_0(x) < \pi_1 f_1(x)$, y en el caso particular de que f_i sea la función de densidad de una normal se tiene que:

$$\pi_0 \cdot |\Sigma_0|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \cdot \exp\left\{-\frac{1}{2}d^2_{M_0}(x,\mu_0)\right\} < \pi_1 \cdot |\Sigma_1|^{-\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \cdot \exp\left\{-\frac{1}{2}d^2_{M_1}(x,\mu_1)\right\}$$

Pasando a dividir $\pi_0, |\Sigma_0|^{-\frac{1}{2}}$ a la derecha, y tomando logaritmos a ambos lados se llega a:

$$-\frac{1}{2}d^2_{M_0}(x,\mu_0) < \log\left(\frac{\pi_1 |\Sigma_1|^{-\frac{1}{2}}}{\pi_0 |\Sigma_0|^{-\frac{1}{2}}}\right) - \frac{1}{2}d^2_{M_1}(x,\mu_1)$$

$$d^2_{M_0}(x,\mu_0) < d^2_{M_1}(x,\mu_1) - 2\log\left(\frac{\pi_1 |\Sigma_1|^{-\frac{1}{2}}}{\pi_0 |\Sigma_0|^{-\frac{1}{2}}}\right)$$

□

Observación: Vemos que esta regla se parece a la regla de Mahalanobis, solo que hay una constante que modifica la inecuación. Si esa constante fuera 0, tendríamos exactamente la regla de Mahalanobis.

¿Cuándo esa constante es 0? Cuando $\Sigma_0 = \Sigma_1$ (es decir, **Homocedasticidad**) y además, cuando $\pi_0 = \pi_1$, es decir, cuando las probabilidades a priori son iguales.

Observación: Ahora, nos venimos arriba y vamos a relacionarlo con la regla de Fisher:

Regla bajo normalidad y homocedasticidad ($\Sigma_0 = \Sigma_1$)
x se clasifica en P_0 si $w'x < w'\left(\frac{\mu_0 + \mu_1}{2}\right) + \log\left(\frac{\pi_0}{\pi_1}\right)$

Esto tiene algún parecido a la regla de Fisher.

6. Regla óptima de clasificación bajo normalidad

- **Homocedasticidad** $\Sigma_1 = \Sigma_0 \rightarrow$ Regla de Fisher trasladada en función de π_0 y π_1.
- Cuando además de homocedasticidad, tenemos $\pi_0 = \pi_1$, la regla óptima es la regla de Fisher tal y como la hemos visto.
- $\Sigma_0 \neq \Sigma_1 \rightarrow$ la distancia de Mahalanobis trasladada.

Error Bayes

Definición 6.1 **Error Bayes**. *El error bayes (L*) es el menor error posible en una clasificación.*

Observación: Dado que la regla de bayes es óptima, esta es una cota inferior para cualquier clasificador.

Apéndice A

Ejercicios

A.1. Hoja 1

Ejercicio 1.1: *Sea* $\mathbf{Y} = (Y_1, Y_2, Y_3)' \equiv N_3(\mu, \Sigma)$*, donde*

$$\mu = (0,0,0)' \; \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

a) *Calcula la distribución del vector* $\mathbf{X} = (X_1, X_2)$*, donde* $X_1 = Y_1 + Y_3$ *y* $X_2 = Y_2 + Y_3$*.*

b) *¿Existe alguna combinación lineal de las variables aleatorias* Y_i *que sea independiente de* X_1*?*

Hecho por Dejuan. Se aceptan correcciones.

APARTADO A)

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} Y_1 + Y_3 \\ Y_2 + Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

Ya tenemos la matriz A que cumple $\mathbf{X} = A\mathbf{Y}$. Utilizando las propiedades de esperanza y varianza (1):

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}(A\mathbf{Y}) = A\mathbb{E}(\mathbf{Y}) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbb{V}(\mathbf{X}) = \mathbb{E}(A\mathbf{Y}) = A\Sigma A' = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

Conclusión:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \equiv N_1 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \right)$$

APARTADO B)

Llamos $Z = a_1Y_1 + a_2Y_2 + a_3Y_3$.

Estas variables serán independientes si se distribuyen conjuntamente como una normal multidimensional y si $\operatorname{cov}(Z, X_1) = 0$.

Vamos a ver la covarianza. Utilizando la propiedad definida en 3, tenemos que

$$\operatorname{cov}(a_1Y_1 + a_2Y_2 + a_3Y_3, X_1) = \operatorname{cov}(A\mathbf{Y}, B\mathbf{Y})$$

Siendo $A = (a_1, a_2, a_3)$ y $B = (1, 0, 1)$

Entonces

$$\operatorname{cov}(A\mathbf{Y}, B\mathbf{Y}) = (a_1, a_2, a_3) \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Operando obtenemos $\operatorname{cov}(A\mathbf{Y}, X_1) = a_1 - a_2 + 2a_3$.

Ahora sólo hace falta ver que se distribuyen conjuntamente como una normal bivariante. Esto lo tenemos asegurado, pues "*El vector se distribuye normalmente porque lo podemos escribir en la forma AY, para una matriz A.*"[1]

Ejercicio 1.2:

Sea $\mathbf{X} = (X_1, X_2, X_3)$ *un vector aleatorio con distribución normal tridimensional con vector de medias* $\mu = (0, 0, 0)$ *y matriz de covarianzas*

$$\Sigma = \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

a) *Determina razonadamente cuáles de los siguientes pares de variables o vectores aleatorios son independientes y cuáles no:*

[1] Cito textualmente de un correo envíado por José Ramón, profesor de la asignatura

***(i)**: X_1 y X_2*

***(ii)**: (X_1, X_3) y X_2*

***(iii)**: X_1 y $X_1 + 3X_2 - 2X_3$*

***b)** Determina una matriz B tal que la variable aleatoria $(X_2, X_3)B(X_2, X_3)'$ tenga distribución χ^2_2.*

APARTADO A)

(i) X_1 y X_2 son independientes porque son marginales de una distribución multivariante conjunta y tienen covarianza 0 (elemento a_{12} de la matriz)

(ii) X_1 y X_2 son independientes porque son marginales de una distribución multivariante conjunta y tienen de matriz de covarianzas el vector idénticamente nulo. Vamos a verlo, aunque para ello construimos $\mathbf{Z} = (X_1, X_3, X_2)$, cuya matriz de covarianzas es:

$$\Sigma_z = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Entonces $\operatorname{cov}\left((X_1, X_3)', X_2\right) = \begin{pmatrix} \operatorname{cov}(X_1, X_2) \\ \operatorname{cov}(X_3, X_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

(iii) X_1 y $X_1 + 3X_2 - 2X_3$. Utilizamos: $\operatorname{cov}(X_1 + 3X_2 - 2X_3, X_1) = \operatorname{cov}(A\mathbf{X}, B\mathbf{X}) = A\Sigma B' = B\Sigma A'$

$$\operatorname{cov}(X_1 + 3X_2 - 2X_3, X_1) = (1, 3, -2) \begin{pmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \ldots = 6$$

Como la covarianza no es cero, entonces existe una relación lineal entre las variables y por ello no son independientes.

APARTADO B)

Una χ^2_k es la distribución que tiene la suma de variables normales estandarizadas al cuadrado. Los k grados de libertad corresponden a la cantidad de variables normales que sumamos.

Vemos que si tomamos $B = I$, obtenemos:

$$(X_2, X_3) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_2 \\ X_3 \end{pmatrix} = X_2^2 + X_3^2$$

Ya tenemos la suma los cuadrados de normales. Ahora sólo falta que estén estandarizadas, es decir que $X_i \sim N(0, 1)$.

Ya están centradas en 0, con lo que sólo falta dividir por la varianza, es decir:

$$(X_2, X_3)\begin{pmatrix}\frac{1}{5} & 0\\ 0 & \frac{1}{2}\end{pmatrix}\begin{pmatrix}X_2\\ X_3\end{pmatrix} = \frac{1}{5}X_2^2 + \frac{1}{2}X_3^2 = Z_2^2 + Z_3^2$$

donde

$$Z_2 = \frac{1}{5}X_2^2 = \left(\frac{X_2}{\sqrt{5}}\right)^2 \to Z_2 \sim N(0,1)$$

$$Z_3 = \frac{1}{2}X_2^2 = \left(\frac{X_2}{\sqrt{2}}\right)^2 \to Z_3 \sim N(0,1)$$

Ejercicio 1.3: *Sea (X, Y) un vector aleatorio con distribución normal bidimensional. Tanto X como Y tienen distribución normal estándar. La covarianza entre X e Y es ρ, donde $|\rho| < 1$.*

a) *Determina cuál es la distribución del vector $(2X - 3Y, X + Y)$.*

b) *Determina cuál es la distribución de la variable $(X^2 - 2\rho XY + Y^2)/(1 - \rho^2)$.*

Hecho por Dejuan. Se aceptan correcciones.

APARTADO A)

Llamamos

$$C = \begin{pmatrix}2 & -3\\ 1 & 1\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$$

Tenemos que calcular $\mathbb{E}(C), \mathbb{V}(C)$. Para ello, utilizamos las fórmulas de siempre

$$\mathbb{E}(C) = \mathbb{E}\left(A\begin{pmatrix}X\\ Y\end{pmatrix}\right) = A\mathbb{E}\left((X,Y)'\right) = A(0,0)' = \begin{pmatrix}0\\ 0\end{pmatrix}$$

$$\mathbb{V}(C) = \mathbb{V}\left(C(X,Y)'\right) = C\Sigma C' = \begin{pmatrix}2 & -3\\ 1 & 1\end{pmatrix}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\begin{pmatrix}2 & 1\\ -3 & 1\end{pmatrix}$$

La distribución del vector $(X, Y) \sim N_2\left(\mathbb{E}(C), \mathbb{V}(C)\right)$

APARTADO B)

Sea

$$Z = \frac{Z_n}{Z_d} = \frac{(X^2 - 2\rho XY + Y^2)}{(1 - \rho^2)}$$

Vemos que

$$Z_n = (X,Y)\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix} = aX^2 + cXY + bXY + dY^2 \implies \begin{cases} a = d = 1 \\ c + b = -2\rho \to c = b = -\rho \end{cases}$$

Ahora, dividimos todo por Z_d. ¿Qué hemos obtenido?

$$\frac{1}{1-\rho^2}\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}$$

Casualmente, esta matriz es la inversa de Σ

$$\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\frac{1}{1-\rho^2}\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

Con lo que

$$Z=(X,Y)\Sigma^{-1}(X,Y)'=(X-0,Y-0)\Sigma^{-1}(X-0,Y-0)'\sim\chi^2_2$$

Ejercicio 1.4: *Sean Y_1 e Y_2 dos variable aleatorias independientes con distribución normal estándar.*

a) *Demuestra que el vector $\mathbf{Y}=(Y_1,Y_2)$ tiene distribución normal bidimensional y calcula la distribución del vector $\mathbf{X}=(2Y_1+Y_2,Y_2-2Y_1)$.*

b) *¿Son las dos distribuciones marginales de X independientes? Determina una matriz B tal que $X'BX$ tenga distribución χ^2 con 2 grados de libertad.*

Hecho por Dejuan. Se aceptan correcciones.

Revisado por Jorge. Se siguen aceptando correcciones

APARTADO A)

Hecho por Jorge. Se aceptan correcciones.

Tomemos la función característica del vector aleatorio que tiene ambas v.a. $Y=(Y_1,Y_2)$:

$$\varphi_Y(t)=\mathbb{E}(e^{it'Y})=\mathbb{E}(e^{it_1Y_1+it_2Y_2})=$$

Puesto que Y_1,Y_2 son independientes:

$$=\mathbb{E}(e^{it_1Y_1})\cdot\mathbb{E}(e^{it_2Y_2})=\varphi_{Y_1}(t_1)\cdot\varphi_{Y_2}(t_2)=e^{-\frac{t_1^2}{2}}\cdot e^{-\frac{t_2^2}{2}}=e^{-\frac{t_1^2+t_2^2}{2}}$$

Que coincide con la función característica de una normal bidimensional $Y\sim N_2(0,I)$.

El vector de n normales independientes se distribuye normalmente. En este caso, como Y_1,Y_2 son normales independientes, $(Y_1,Y_2)\sim N(\mu,\Sigma)$, donde:

$$\mu=\begin{pmatrix}0\\0\end{pmatrix}\quad\Sigma=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$\mathbf{X} = (2Y_1 + Y_2, Y_2 - 2Y_1) \rightarrow \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

Entonces, vamos a calcular la distribución de $\mathbf{X}$

$$\mathbb{E}(\mathbf{X}) = \mathbb{E}(A\mathbf{Y}) = A\mathbb{E}(Y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbb{V}(\mathbf{X}) = \mathbb{V}(A\mathbf{Y}) = A\mathbb{V}(\mathbf{Y})A' = AA' = AA = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$$

APARTADO B)

$X_i \sim N(0,5)$. Además, $\text{corr}(X_1, X_2) \neq 0$.

Por tanto no son independientes debido a que la correlación entre ambas no es cero.

Sabemos que una χ^2_2 es la suma de dos normales estandarizadas al cuadrado $\Sigma^{-1/2}(X - \mu) = Y \sim N_2(0, I)$:

$$\chi^2_2 = Y_1^2 + Y_2^2 = Y'Y = (X-\mu)'\Sigma^{-1/2}\Sigma^{-1/2}(X-\mu) \overset{\mu=0}{=} X'\Sigma^{-1}X$$

Por tanto la B que pide el enunciado no es más que:

$$\begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}^{-1}$$

Ejercicio 1.5:

Sea (X, Y) un vector aleatorio con función de densidad

$$f(x,y) = \frac{1}{2\pi}\exp\left[\frac{1}{2}\left(x^2 - 2xy + 2y^2\right)\right]$$

***a)** Calcula la distribución condicionada de X dado $Y = y$, y la de Y dado $X = x$.*

Mirando la función de densidad y comparándola con la de la normal, podemos escribir:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \equiv N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}^{-1}\right) \equiv N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$$

Aplicando las fórmulas vistas en teoría 1.10, nos damos cuenta de que tenemos que calcular $X_2|X_1$ y $X_1|X_2$, con lo que cada caso tendrá una pequeña variación en la fórmula:

$$E(X|Y=y) = \mu_y + \Sigma_{12}\Sigma_{22}^{-1}(X-\mu_x) = 0 + \frac{1}{1}(y-0) = y$$
$$E(Y|X=x) = \mu_x + \Sigma_{21}\Sigma_{11}^{-1}(Y-\mu_y) = 0 + \frac{1}{2}(x-0) = \frac{x}{2}$$

Ejercicio 1.6: *Sea* $\mathbf{X} = (X_1, X_2)$ *un vector aleatorio con distribución normal bidimensional con vector de medias* $(1,1)$ *y matriz de covarianzas*

$$\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$$

Calcula la distribución de $X_1 + X_2$ *condicionada por el valor de* $X_1 - X_2$ *.*

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

Entonces, calculando como siempre obtenemos:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \equiv N_2 \left(\begin{pmatrix} 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix} \right)$$

Sabemos que la distribución va a ser normal, por lo que necesitamos $\mathbb{E}\left(Z_1|Z_2\right)$ y $\mathbb{V}\left(Z_1|Z_2\right)$

Utilizando las fórmulas tenemos:

$$\mathbb{E}\left(Z_1|Z_2\right) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \mu_2) = 2 + 1\frac{1}{3}(Z_2 - 0) = \frac{7}{3}Z_2$$

$$\mathbb{V}\left(Z_1|Z_2\right) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 7 - 1\frac{1}{3}1 = \frac{20}{3}$$

Entonces,

$$(Z_2|Z_1) = (X_1 + X_2|X_1 - X_2) \sim N_2\left(\frac{7}{3}(X_1 - X_2), \frac{20}{3}\right)$$

Ejercicio 1.7: *Sea* $X = (X1, X2, X3)'$ *un vector aleatorio con distribución normal tridimensional con vector de medias* $(0,0,0)'$ *y matriz de covarianzas*

$$\Sigma = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 4 \end{pmatrix}$$

Definamos las v.a. $Y_1 = X_1 + X_3, Y_2 = 2X_1 - X_2$ *e* $Y_3 = 2X_3 - X_2$. *Calcula la distribución de* Y_3 *dado que* $Y_1 = 0$ *e* $Y_2 = 1$.

Lo primero es descubrir la matriz de la combinación lineal y calcular la distribución, esto es:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \equiv N_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 & -2 & 4 \\ -2 & 2 & -2 \\ 4 & -2 & 22 \end{pmatrix} \right)$$

Ahora vamos a calcular las condicionadas. Sabemos que $Y_3|Y_1 = 0, Y_2 = 1 \sim N_1(\mu_{2.1}, \Sigma_{2.1})$.

Hacemos la división:

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right) = \left(\begin{array}{cc|c} 3 & -2 & 4 \\ -2 & 2 & -2 \\ \hline 4 & -2 & 22 \end{array} \right)$$

$$E(Y_3|Y_1 = 0, Y_2 = 1) = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1} \begin{pmatrix} Y_1 - \mu_1 \\ Y_2 - \mu_2 \end{pmatrix} = 0 + (4, -2) \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 - 0 \\ 1 - 0 \end{pmatrix}$$

$$V(Y_3|Y_1 = 0, Y_2 = 1) = \Sigma_2 2 - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 22 - (4, -2) \begin{pmatrix} 3 & -2 \\ -2 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

Terminando las cuentas: $E(Y_3|Y_1 = 0, Y_2 = 1) = 1$ y $V(Y_3|Y_1 = 0, Y_2 = 1) = 16$.

Entonces, la distribución de $(Y_3|Y_1 = 0, Y_2 = 1) = N_1(1, 16)$

Ejercicio 1.8: *Sea* $Y = (Y_1, ..., Y_n)$ *un vector normal multivariante tal que las coordenadas* Y_i *tienen distribución* $N(0, 1)$ *y, además,* $\operatorname{cov}(Y_i, Y_j) = \rho$, *si* $i \neq j$.

a) *Escribe el vector de medias y la matriz de covarianzas del vector* $X = (Y_1 + Y_2, Y_1 - Y_2)'$ *. ¿Son* $Y_1 + Y_2$ *e* $Y_1 - Y_2$ *dos variables aleatorias independientes?*

b) *Si* Σ *es la matriz de covarianzas de* X, *¿cuál es la distribución de la variable aleatoria* $Z = X'\Sigma^{-1}X$*?*

c) *Si* $\rho = 1/2$, *calcula la varianza de la media muestral* $\overline{Y} = (Y_1 + \cdots + Y_n)/n$ *(en función del tamaño muestral* n*).*

Hecho por Dejuan. Se aceptan correcciones.

Revisado por Jorge. Se siguen aceptando correcciones

APARTADO A)

Tenemos:

$$X = \begin{pmatrix} Y_1 + Y_2 \\ Y_1 - Y_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & -1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

El vector de medias es $\mu = \mathbb{E}(A\mathbf{Y}) = A\mathbb{E}(\mathbf{Y}) = (0,0)'$

La matriz de covarianzas:

$$\mathbb{V}(A\mathbf{Y}) = A\mathbb{V}(\mathbf{Y})A' = \dots = \begin{pmatrix} 2+2\rho & 0 \\ 0 & 2-2\rho \end{pmatrix}$$

Como $\operatorname{corr}(X_1, X_2) = 0$ y ambas variables vienen de un vector normal, concluimos que son independientes.

Otra manera mucho más corta es utilizar la 1.12.

En este caso, $A = (1,1,0,...,0)$ y $B = (1,-1,0,...,0)$. Como $AB' = 0 \implies AY = (Y_1 + Y_2)$ y $BY = (Y_1 - Y_2)$ son independientes.

¿Boom?

APARTADO B)

Una χ^2_2 ya que estamos sumando 2 variables normales estandarizadas (se estandarizan al tener la forma cuadrática Σ^{-1} y tener vector de medias nulo).

APARTADO C)

Tenemos la matriz de combinación lineal $A = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$. Como sólo nos piden la varianza:

$$\mathbb{V}(A\mathbf{Y}) = A\mathbb{V}(\mathbf{Y})A' = \frac{1}{n^2} 1_n \Sigma 1_n' =$$

$$\frac{1}{n^2}(1,1,...,1) \begin{pmatrix} 1 & \frac{1}{2} & \dots & \dots & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} & \dots & \frac{1}{2} \\ \vdots & & \ddots & & \vdots \\ \frac{1}{2} & & \dots & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \dots = \frac{1}{n^2} \frac{n(n+1)}{2} = \frac{n+1}{2n}$$

$$\mathbb{V}\left(\overline{Y}\right) = \mathbb{V}(A\mathbf{Y}) = \frac{n+1}{2n}$$

Ejercicio 1.9: *Demuestra que si X es un vector aleatorio con distribución $N_k(\mu, \Sigma)$, entonces existen $\lambda_1, ..., \lambda_k \in \mathbb{R}^+$ y v.a.i.i.d. $Y_1, ..., Y_k$ con distribución χ^2_1 tales que $||X - \mu||^2$ se distribuye igual que $\lambda_1 Y_1 + \cdots + \lambda_k Y_k$.*

En particular, deduce que si Σ es simétrica e idempotente y $\mu = 0$, entonces $||X||^2$ tiene distribución χ^2_r donde r es la traza de Σ

Sabemos que $\Sigma = CDC'$ con C una matriz formada por autovectores ortonormales. Puesto que $X - \mu \sim N(0, \Sigma)$ PODEMOS continuar de la siguiente forma:

$$Z = C'(X - \mu) \sim N_k(0, D)$$

$$\|X - \mu\|^2 = (X - \mu)'(X - \mu) = Z' \underbrace{C'C}_{I} Z = Z'Z = \sum_{i=1}^{k} Z_i^2$$

Ya que $Z_i \sim N(0, \lambda_i)$ con λ_i el elemento i-ésimo de la matriz diagonal D, se tiene que:

$$Y_i = \frac{Z_i^2}{\lambda_i} \sim \chi^2_1$$

Y por tanto $\sum_{i=1}^k Z_i^2 = \sum_{i=1}^k \lambda_i Y_i$

En el caso particular de que Σ sea simétrica e idempotente, sus autovalores son $\lambda_i = 0, 1$, de modo que se pasa a tener (con $\mu = 0$):

$$\|X\|^2 = \sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{r} Y_i \sim \chi^2_r$$

Donde r es el número de autovalores $\lambda_i = 1$ de D, dicho número coincide precisamente con el rango de Σ.

A.2. Hoja 2

Ejercicio 2.1: *Calcula la distribución exacta bajo la hipótesis nula del estadístico de Kolmogorov-Smirnov para muestras de tamaño 1.*

La hipótesis sería $H_0 : F = F_0$ continua, con $X \sim F$

En este caso,

$$D = ||F_1 - F_0||_{\text{ínf}} = (1) = \text{máx}\{F_0(x), 1 - F_0(x)\}$$

(1) hay 2 posibles caminos. Al dibujar lo que nos dicen (una muestra de tamaño 1) podemos sacarlo por intuición. Sino, aplicamos la fórmula de los estadísticos.

Ahora calculamos:

$$P_{F_0}(D \leq x) = P_{F_0} = \left\{\text{máx}\{...\} \leq x\right\} = P_{F_0} = P_{F_0}\{1 - x \leq F_0(x) \leq x\}$$

No entiendo porqué $P_{F_0}\left\{\text{máx}\{...\} \leq x\right\} = \{1 - x \leq F_0(x) \leq x\}$ y no es $\{x \leq F_0(x) \leq 1 - x\}$

Resolvemos la desigualdad, aplicando que F_0 es una uniforme.

$$P\{1 - x \leq U \leq x\} = \begin{cases} 0 & x \leq \frac{1}{2} \\ 2x - 1 & x \geq \frac{1}{2} \end{cases} \implies D \sim \mathcal{U}\left(\frac{1}{2}, 1\right)$$

Ya que $1 - x > x \iff x \leq \frac{1}{2}$

Ejercicio 2.2: *Se desea contrastar la hipótesis nula de que una única observación X procede de una distribución N(0,1). Si se utiliza para ello el contraste de Kolmogorov-Smirnov, determina para qué valores de X se rechaza la hipótesis nula a nivel α = 0,05.*

Este ejercicio está muy relacionado con el primero. Es una aplicación al caso de la normal.

Mirando en la tabla, encontramos que para $\alpha = 0.05$, entonces $d_\alpha = 0.975$. Con esta inormación podemos construir la región crítica:

$$R = \left\{\text{máx}\{\Phi(x), 1 - \Phi(x))\} > 0.975\right\} = \{\Phi(x) > 0.975\} \cup \{1 - \Phi(x) > 0.975\} =$$

$$\{X > \Phi^{-1}(0.975)\} \cup \{X < \Phi^{-1}(0.025)\}$$

Consultando las tablas, vemos que $\Phi^{-1}(0.025) = 1.96$ y por simetría, $\Phi^{-1}(0.975) = -1.96$

$$R = \{|X| > 1.96\}$$

Observación: Es interesante saber que, al ser simétrica la normal, la interpretación gráfica es muy fácil. Si dividimos la normal en 3 intervalos,

$$(-\infty, -1.96), (-1.96, 1.96), (1.96, \infty)$$

, el área encerrada en las colas es el nivel de significación, en este caso:

$$\text{Area } \left((-\infty, -1.96) \cup (1.96, \infty)\right) = 0.05$$

Ejercicio 2.3: *Da una demostración directa para el caso k = 2 de que la distribución del estadístico del contrast χ^2 de bondad de ajuste converge a una distribución χ^2_1 , es decir,*

$$T = \frac{(O1 - E1)^2}{E1} + \frac{(O2 - E2)^2}{E2} \xrightarrow[n\to\infty]{d} \chi^2_1$$

[Indicación: Hay que demostrar que $T = X_n^2$, donde $X_n \xrightarrow[n\to\infty]{d} N(0, 1)$. Para reducir los dos sumandos a uno, utilizar la relación existente entre O1, E1 y O2, E2.]

Si tenemos n datos, vamos a construir la tabla de contingencia. Creo que consideramos una binomial porque, al sólo tener 2 clases, o eres de una o eres de la otra con una probabilidad p.

	A_1	A_2
Obs	$n\overline{p}$	$n(1-\overline{p})$
Esp	np_0	$n(1-p_0)$

$$T = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{n^2(\overline{p}-p_0)^2}{n} + \frac{n^2(\overline{p}-p_0)}{n(1-p_0)} = ...$$

Simplificando, llegamos a:

$$T = \left(\frac{|\overline{p}-p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right)$$

Está contando un montón de cosas interesantes que me estoy perdiendo.

Entre ellas, tenemos que $\sqrt{T} \xrightarrow[n\to\infty]{d} N(0.1)$ por el teorema central del límite (es el caso particular para una binomial), con lo que $T \xrightarrow[n\to\infty]{d} \chi^2$. ¿Porqué 1 grado de libertad? Porque sólo estamos estimando 1 parámetro, el $\overline{p}$.

Esto responde también al problema 11.

Ejercicio 2.4: *El número de asesinatos cometidos en Nueva Jersey cada día de la semana durante el año 2003 se muestra en la tabla siguiente:*

Día	*Lunes*	*Martes*	*Miércoles*	*Jueves*	*Viernes*	*Sábado*	*Domingo*
Frecuencia	*42*	*51*	*45*	*36*	*37*	*65*	*53*

a) *Contrasta a nivel α = 0,05, mediante un test $\chi 2$, la hipótesis nula de que la probabilidad de que se cometa un asesinato es la misma todos los días de la semana.*

b) *¿Podría utilizarse el test de Kolmogorov-Smirnov para contrastar la misma hipótesis? Si tu respuesta es afirmativa, explica cómo. Si es negativa, explica la razón.*

c) *Contrasta la hipótesis nula de que la probabilidad de que se cometa un asesinato es la misma desde el lunes hasta el viernes, y también es la misma los dos días del fin de semana (pero no es necesariamente igual en fin de semana que de lunes a viernes).*

APARTADO A)

Tenemos $n = 329$, $E_i = \frac{329}{7} = 47$ y $H_0 : p_i = \frac{1}{7}$

Calculamos el estadístico

$$T = \sum_{i=1}^{7} \frac{O_i^2}{E_i} - 329 = \left(\frac{42^2}{47} + \frac{51^2}{47} + \frac{45^2}{47} + ... + \frac{53^2}{47} \right) - 329 = 13.32$$

Por otro lado, $\chi^2_{6;0.05} = 12.59$, con lo que rechazamos la hipótesis.

APARTADO B)

No podría utilizarse al tratarse de algo discreto y KS sólo sirve para continuas.

APARTADO C)

Tenemos la siguiente tabla:

Día	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo
Frecuencia	p	p	p	p	p	q	q

Observación: Podríamos plantearnos contrastar que es uniforme de lunes a viernes (H_1) y otra uniforme distinta en fines de semana (H_2). Entonces tendríamos $H_0 : H_1 \cap H_2$, y construir la región $R = R_1 \cup R_2$. ¿Cuál es el problema de este camino?

El nivel de significación, ya que $P_{H_0}(R_1 \cup R_2) = P_{H_0}(R_1) + P_{H_0}(R_2) - P_{H_0}(R_1 \cap R_2) = 2\alpha - \alpha^2 \sim 2\alpha$.

Podríamos tomar, chapucerillamente $\alpha = \frac{\alpha}{2}$ para que al final, $P_{H_0}(R_1 \cup R_2) = \alpha$. Aquí surge otro problema, que es que estamos despreciando la probabilidad de la intersección y tomándolo como independiente cuando no tiene porqué serlo. Es una aproximación "buena" que a veces se utiliza, pero pudiendo hacerlo bien...

Vamos a hacerlo bien: Tenemos que $5p + 2q = 1 \implies q = \frac{1-5p}{2}$. Pero para utilizar el contraste de homogeneidad χ^2 necesitamos tener p (y q). Como no disponemos de ellos, vamos a estimarlos. ¿Cómo? Con el estimador de máxima verosimilitud que es el molón. En el apéndice hay un pequeño recordatorio: sección B.1

En este caso, nuestra función de densidad es:

$$f(x) = \begin{cases} p & x \in \text{[lunes,martes,miércoles,jueves,viernes]} \\ \frac{1-5p}{2} & x \in \text{[sábado,domingo]} \end{cases}$$

¿Cuál es la probabilidad de 7 asesinatos entre semana? Pues la intersección de los 7 sucesos, es decir $p \cdot p \cdot ... \cdot p = p^7$. Razonando así, tenemos

$$e.m.v.(p) = L(p; \text{datos}) = p^{42+51+...+37} \left(\frac{1-5p}{2}\right)^{65+53}$$

Ahora, despejamos tomando $l(p) = \ln(L(p)) = 211 \ln(p) + 118 \ln\left(\frac{1-5p}{2}\right)$ y maximizamos:

$$l'(p) = 0 \implies ... \begin{cases} \overline{p} = 0.128 \\ \overline{q} = 0.179 \end{cases}$$

Ahora que ya tenemos p y q, las frecuencias esperadas son:

$$E_i = n \cdot (p, p, p, p, p, q, q) = (42.2, ..., 42.2, 58.91, 58.91)$$

Ya estamos en condiciones de construir el estadístico:

$$T = \sum_{i=1}^{7} \frac{O_i^2}{E_i^2} - n = ... = 5.4628$$

Y comparamos con la χ^2. ¿Cuántos grados de libertad? Si tenemos 7 clases, siempre perdemos uno, con lo que serían 6. Sin embargo hemos estimado un parámetro, con lo que son 5 grados de libertad. Entonces: $c = \chi^2_{5;0.05} = 11.07$

Como $T < c$, no podemos rechazar la hipótesis.

Ejercicio 2.5: *Para estudiar el número de ejemplares de cierta especie en peligro de extinción que viven en un bosque, se divide el mapa del bosque en nueve zonas y se cuenta el número de ejemplares de cada zona. Se observa que 60 ejemplares viven en el bosque repartidos en las 9 zonas de la siguiente forma:*

8	*7*	*3*
5	*9*	*11*
6	*4*	*7*

Mediante un contraste de hipótesis, analiza si estos datos aportan evidencia empírica de que los animales tienen tendencia a ocupar unas zonas del bosque más que otras.

Tomamos $\alpha = 0.01$

$T = 7.47$, $\chi^2_{8;0.001} = 20.09$

Aceptamos la hipótesis H_0 : la especie se reparte uniformemente.

Ejercicio 2.6: *Se ha desarrollado un modelo teórico para las diferentes clases de una variedad de moscas. Este modelo nos dice que la mosca puede ser de tipo L con probabilidad p 2 , de tipo M con probabilidad q 2 y de tipo N con probabilidad 2pq (p + q = 1). Para confirmar el modelo experimentalmente tomamos una muestra de 100 moscas, obteniendo 10, 50 y 40, respectivamente.*

***a)** Hallar la estimación de máxima verosimilitud de p con los datos obtenidos.*

***b)** ¿Se ajustan los datos al modelo teórico, al nivel de significación 0'05?*

Hecho por Jorge. Se aceptan correcciones.

Revisado por Dejuan. Se siguen aceptando correcciones

APARTADO A)

Primero calculamos la función de verosimilitud para p:

$$L_n(p) = L_n(p) = \prod_{i=0}^{n} f(x_i; p) = (p^2)^{10} \cdot (q^2)^{50} \cdot (2pq)^{40}$$

El EMV lo obtendremos maximizando $\log L_n(p)$:

$$\log L_n(p) = 20 \log p + 100 \log q + 40 \log 2pq$$

$$\frac{\partial}{\partial p} \log L_n(p) = \frac{20}{p} - \frac{100}{1-p} + 40 \frac{2-4p}{2p(1-p)} = 0$$

Maximizamos con $\hat{p} = \frac{3}{10} \implies \hat{q} = \frac{7}{10}$.

APARTADO B)

En este caso tomamos $H_0 \equiv P(X \in L) = p^2, P(X \in M) = q^2, P(X \in N) = 2pq$

Usando el estado el contraste de bondad de ajuste de la χ^2, el estadístico de Pearson queda:

$$T = \sum_{i=1}^{3} \frac{\left(O_i - \hat{E}_i\right)^2}{\hat{E}_i} = \sum_{i=1}^{3} \frac{O_i^2}{\hat{E}_i} - n =$$

$$= \frac{10^2}{p^2 \cdot 100} + \frac{50^2}{(1-p)^2 \cdot 100} + \frac{40^2}{2p(1-p) \cdot 100} - 100 \approx 0.22$$

Puesto que en este caso $k = 3$ y hemos estimado 1 parámetro (p), tenemos que T se distribuye como una χ^2_{3-1-1}. En las tablas nos encontramos con que $\chi^2_{1;0.05} = 3.84 > T$ y no rechazamos H_0, es decir los datos se ajustan al modelo teórico.

Ejercicio 2.7:

a) *Aplica el test de Kolmogorov-Smirnov, al nivel 0.05, para contrastar si la muestra (3.5, 4, 5, 5.2, 6) procede de la $U(3, 8)$.*

b) *Aplica el test de Kolmogorov-Smirnov, al nivel 0.05, para contrastar la hipótesis de que la muestra (0, 1.2, 3.6) procede de la distribución $N(\mu = 1; \sigma = 5)$.*

Hecho por Jorge. Se aceptan correcciones.

Revisado por Dejuan. Se siguen aceptando correcciones

APARTADO A)

La función de distribución de una $U(3, 8)$ es:

$$F(x) = \begin{cases} 0 & , x < 3 \\ \frac{x-3}{5} & , 3 \leq x \leq 8 \\ 1 & , x > 8 \end{cases}$$

$x_{(i)}$	$\frac{i}{n}$	$F_0(x_{(i)})$	D_n^+	D_n^-
3.5	0.2	0.1	0.1	0.1
4	0.4	0.2	0.2	0
5	0.6	0.4	0.2	0
5.2	0.8	0.44	0.36	-0.16
6	1	0.6	0.4	-0.2

Tendremos por tanto que $D_n = \|F_n - F_0\|_\infty = 0.4$. Si nos vamos a la tabla del contraste K-S vemos que $c = 0.565$ para $\alpha = 0.05$.

Como $D_n < c$ **no rechazamos** la hipótesis nula de que las muestras vienen de la uniforme.

APARTADO B)

$x_{(i)}$	$\frac{i}{n}$	$F_0(x_{(i)})$	D_n^+	D_n^-
0	0.3	0.42	-0.12	0.42
1.2	0.6	0.52	0.08	0.22
3.6	1	0.7	0.3	0.1

Tendremos por tanto que $D_n = \|F_n - F_0\|_\infty = 0.42$. Si nos vamos a la tabla del contraste K-S vemos que $c = 0.708$ para $\alpha = 0.05$.

Como $D_n < c$ **no rechazamos** la hipótesis nula de que las muestras vienen de la $N(1, 5)$.

Ejercicio 2.8: *Se ha clasificado una muestra aleatoria de 500 hogares de acuerdo con su situación en la ciudad (Sur o Norte) y su nivel de renta (en miles de euros) con los siguientes resultados:*

Renta	*Sur*	*Norte*
0 a 10	*42*	*53*
10 a 20	*55*	*90*
20 a 30	*47*	*88*
más de 30	*36*	*89*

a) *A partir de los datos anteriores, contrasta a nivel α = 0,05 la hipótesis nula de que en el sur los hogares se distribuyen uniformemente en los cuatro intervalos de renta considerados.*

b) *A partir de los datos anteriores, ¿podemos afirmar a nivel α = 0,05 que la renta de los hogares es independiente de su situación en la ciudad?*

APARTADO A)

Tenemos $H_0 : p_i = \frac{1}{4}$ y usando el contraste de bondad de ajuste de la χ^2:

$$T = \sum_{i=1}^{4} \frac{O_i^2}{E_i} - n_{\text{sur}} = \frac{42^2 + 55^2 + 47^2 + 36^2}{\frac{1}{4} \cdot 180} - 180 = 4.31$$

En las tablas encontramos que $\chi^2_{k-1;\alpha} = \chi^2_{3;0.05} = 7.815$. Como $T < \chi^2_{3;0.05}$, **no podemos rechazar** la hipótesis nula de que en el sur los hogares se distribuyen uniformemente en los cuatro intervalos de renta considerados.

APARTADO B)

Lo primero que haremos es estimar las probabilidades de que la v.a. caiga en cada una de las 6 clases que tenemos (A_i serán los intervalos de renta y B_i si el hogar es del norte o del sur):

$$p(x \in A_1) = \frac{42 + 53}{500} = 0.19$$
$$p(x \in A_2) = \frac{55 + 90}{500} = 0.29$$
$$p(x \in A_3) = \frac{47 + 88}{500} = 0.27$$
$$p(x \in A_4) = \frac{36 + 89}{500} = 0.25$$

$$p(x \in B_1) = \frac{42 + 55 + 47 + 36}{500} = 0.36$$
$$p(x \in B_2) = \frac{53 + 90 + 88 + 89}{500} = 0.64$$

Bajo la H_0 consideramos A_i independiente de B_i, de modo que $p_{i,j} = p_i \cdot p_j$ tal y como se muestra en la siguiente tabla:

$p_{1,1} = 0.0684$	$p_{1,2} = 0.1216$
$p_{2,1} = 0.1044$	$p_{2,2} = 0.1856$
$p_{3,1} = 0.0972$	$p_{3,2} = 0.1728$
$p_{4,1} = 0.09$	$p_{4,2} = 0.16$

Sabiendo que $\hat{E}_{ij} = n \cdot p_{i,j}$:

$\hat{E}_{1,1} = 34.2$	$\hat{E}_{1,2} = 60.8$
$\hat{E}_{2,1} = 52.2$	$\hat{E}_{2,2} = 92.8$
$\hat{E}_{3,1} = 48.6$	$\hat{E}_{3,2} = 86.4$
$\hat{E}_{4,1} = 45$	$\hat{E}_{4,2} = 80$

$$T = \sum_{j=1}^{2}\sum_{i=1}^{4}\frac{O_{ij}^2}{\hat{E}_{ij}} - n = 8.39$$

Si nos vamos a las tablas vemos que $\chi^2_{(k-1)(p-1);\alpha} = \chi^2_{3\cdot 1;0.05} = 7.815 < T$ y por tanto **rechazamos la hipótesis nula** de que la renta de los hogares es independiente de su situación en la ciudad.

Hecho por Dejuan. Se aceptan correcciones.

$$T = \sum_{j=1}^{2}\sum_{i=1}^{4}\frac{O_{ij}^2}{\hat{E}_{ij}} - n = 5.91 < 7.815$$

y por tanto **aceptamos la hipótesis nula** de que la renta de los hogares es independiente de su situación en la ciudad.

Ejercicio 2.9: *A finales del siglo XIX el físico norteamericano Newbold descubrió que la proporción de datos que empiezan por una cifra d, p(d), en listas de datos correspondientes a muchos fenómenos naturales y demográficos es aproximadamente: p(d) = log10 d + 1 d ! , d = 1,2,...,9. Por ejemplo, p(1) = log10 2 ≈ 0,301030 es la frecuencia relativa de datos que empiezan por 1. A raíz de un artículo publicado en 1938 por Benford, la fórmula anterior se conoce como ley de Benford. El fichero poblacion.RData incluye un fichero llamado poblaciones con la población total de los municipios españoles, así como su población de hombres y de mujeres. (a) Contrasta a nivel α = 0,05 la hipótesis nula de que la población total se ajusta a la ley de Benford. (b) Repite el ejercicio pero considerando sólo los municipios de más de 1000 habitantes. (c) Considera las poblaciones totales (de los municipios con 10 o más habitantes) y contrasta a nivel α = 0,05 la hipótesis nula de que el primer dígito es independiente del segundo. (Indicación: Puedes utilizar, si te sirven de ayuda, las funciones del fichero benford.R).*

Ejercicio 2.10: *Se ha llevado a cabo una encuesta a 100 hombres y 100 mujeres sobre su intención de voto. De las 100 mujeres, 34 quieren votar al partido A y 66 al partido B. De los 100 hombres, 50 quieren votar al partido A y 50 al partido B.*

a) *Utiliza un contraste basado en la distribución χ^2 para determinar si con estos datos se puede afirmar a nivel $\alpha = 0,05$ que el sexo es independiente de la intención de voto.*

b) *Determina el intervalo de valores de α para los que la hipótesis de independencia se puede rechazar con el contraste del apartado anterior.*

Este ejercicio ha caido en un examen.

Hecho por Jorge. Se aceptan correcciones.

Revisado por Dejuan. Se siguen aceptando correcciones

APARTADO A)

Procediendo como en el ejercicio anterior obtendremos que bajo la hipótesis nula de independencia:

$$p_{A,\text{mujer}} = p_{A,\text{hombre}} = 0.21$$
$$p_{B,\text{mujer}} = p_{B,\text{hombre}} = 0.29$$

Por tanto:

$$T = \sum_{j=1}^{2}\sum_{i=1}^{2}\frac{O_{ij}^2}{\hat{E}_{ij}} - 200 = 5.25$$

Si nos vamos a las tablas vemos que $\chi^2_{(k-1)(p-1);\alpha} = \chi^2_{1;0.05} = 3.841 < T$, y por tanto **rechazamos la hipótesis nula** de que el sexo es independiente de la intención de voto.

En clase: hemos contrastado homogeneidad (las intenciones de voto se distribuyen igual) en vez de independencia, pero viene a ser lo mismo.

APARTADO B)

El p-valor asociado a $T = 5.25$ es $\left[1 - F_{\chi^2_1}(5.25)\right] = 0.02$, por tanto para $\alpha \in [0.02, 1]$ rechazamos la hipótesis de independencia del apartado anterior.

Para calcular el p-valor, utilizamos que una χ^2_1 es una normal al cuadrado, es decir:

$$p = P(X > 5.25) = P(Z^2 > 5.25) = P(|Z| > 2.29) = 0.022$$

siendo $Z \sim N(0, 1)$

Ejercicio 2.11: *Sea X1,...,Xn una muestra de una distribución Bin(1, p). Se desea contrastar H0 : p = p0. Para ello hay dos posibilidades:*

a) *Un contraste de proporciones basado en la región crítica $R = \{|\overline{p} - p_0|\} > z\frac{\alpha}{2}pp0(1-p0)/n$*

b) *un contraste $\chi 2$ de bondad de ajuste con k = 2 clases. ¿Cuál es la relación entre ambos contrastes?*

Consultar el ejercicio 2.3.

Ejercicio 2.12: *En un estudio de simulación se han generado 10000 muestras aleatorias de tamaño 10 de una distribución $N(0,1)$. Para cada una de ellas se ha calculado con R el estadístico de Kolmogorov-Smirnov para contrastar la hipótesis nula de que los datos proceden de una distribución normal estándar, y el correspondiente p-valor.*

a) *Determina un valor x tal que la proporción de estadísticos de Kolmogorov-Smirnov mayores que x, entre los 10000 obtenidos, sea aproximadamente igual a 0.05. ¿Cuál es el valor teórico al que se debe aproximar la proporción de p-valores menores que 0.1 entre los 10000 p-valores obtenidos?*

b) *¿Cómo cambian los resultados del apartado anterior si en lugar de considerar la distribución normal estándar se considera una distribución uniforme en el intervalo (0,1)?*

Hecho por Jorge. Se aceptan correcciones.

APARTADO A)

- La x que nos piden es $f_{D,\alpha=0.05}$ (f_D es la función de densidad del estadístico K-S). Si acudimos a la tabla vemos que para $n = 10$ $x = f_{D,0.05} = 0.41$. Un poco más explicado el razonamiento:

$$\underbrace{\frac{\#\{i : D_i > x\}}{10000}}_{P(D>x)} \simeq 0.05$$

- Precisamente el $10\,\%$ de los p-valores debería ser menor que 0.1, ya que hacer un contraste nivel de significación $\alpha = 0.1$ significa que en el $10\,\%$ de los casos rechazamos la hipótesis nula, es decir, en le $10\,\%$ de los casos los p-valores son < 0.1.

  Esto se debe al concepto de nivel de significación, ya que si el nivel de significación es 0.01, entonces nos estamos equivocando en 1 de cada 100 contrastes que hagamos, es decir:

$$\frac{\#\{i : p^{(i)} < \alpha\}}{10000} \simeq \alpha$$

APARTADO B)

- Al contrastar con una distribución $U(0,1)$ cabría esperar que las 1000 D_i tomaran valores más altos, pues la distancia entre F_n (que se monta a partir de datos que vienen de una $N(0,1)$) y $F_0 = F_{U(0,1)}$ sería más grande que al tomar como F_0 la de una $N(0,1)$. Por tanto **el valor x debería ser mayor**.
- Por otra parte la proporción de p-valores menores que 0.1 debería aumentar, ya que el test debería devolver p-valores más pequeños (pues debería de rechazar la hipótesis de que los datos vienen de una $U(0,1)$).

Solución de clase: Al tener muchas muchas muestras, las frecuencias deberían ser las probabilidades.

A.3. Hoja 3

Ejercicio 3.1: *La Comunidad de Madrid evalúa anualmente a los alumnos de sexto de primaria de todos los colegios sobre varias materias. Con las notas obtenidas por los colegios en los años 2009 y 2010 (fuente: diario El País) se ha ajustado el modelo de regresión simple:*

$$Nota2010 = \beta_0 + \beta_1 Nota2009 + \varepsilon,$$

en el que se supone que la variable de error ε verifica las hipótesis habituales. Los resultados obtenidos con R fueron los siguientes:

Coefficients:

	Estimate	*Std. Error*	*t-value*	*Pr(>|t|)*
(Intercept)	*1.40698*	*0.18832*	*7.471*	*1.51e-13*
nota09	*0.61060*	*0.02817*	*21.676*	*< 2e-16*

Residual standard error: 1.016 on 1220 degrees of freedom

Multiple R-squared: 0.278,Adjusted R-squared: 0.2774

F-statistic: 469.8 on 1 and 1220 DF, p-value: < 2.2e-16

También se sabe que en 2009 la nota media de todos los colegios fue 6,60 y la cuasidesviación típica fue 1,03 mientras que en 2010 la media y la cuasidesviación típica fueron 5,44 y 1,19, respectivamente.

a) *¿Se puede afirmar a nivel $\alpha = 0,05$ que existe relación lineal entre la nota de 2009 y la de 2010? Calcula el coeficiente de correlación lineal entre las notas de ambos años.*

b) *Calcula un intervalo de confianza de nivel $95\,\%$ para el parámetro β_1 del modelo.*

c) *Calcula, a partir de los datos anteriores, un intervalo de confianza de nivel $95\,\%$ para la nota media en 2010 de los colegios que obtuvieron un 7 en 2009.*

Hecho por Jorge. Se aceptan correcciones.

Revisado por Dejuan. Se siguen aceptando correcciones

APARTADO A)

Poniendo $H_0 : \beta_1 = 0$ (no hay relación lineal entre las notas de uno y otro año) tendremos:

$$\frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} \equiv t_{n-2}$$

La salida nos dice que este estadístico sale 21.676, y el p-valor asociado es $< 2e-16 < 0.05 = \alpha$. Por tanto rechazamos la hipótesis nula H_0, y podemos afirmar que existe relación lineal entre la nota de 2009 y la de 2010.

Jorge: no lo tengo muy claro, pero creo que la segunda pregunta de este apartado pide $\hat{\beta}_1$. Y según la salida de R eso es 0.61

APARTADO B)

La definición del intervalo de confianza de nivel $95\,\%$ para β_1 es:

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 \mp t_{n-2;\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}}\right] \overbrace{=}^{\text{salida R}} \left[0.61 \mp t_{1220;0.025} \cdot 0.02\right]$$

Si buscamos en las tablas de la t, no encontramos para más grados de libertad que 100. ¿Por qué? Porque una t con tantos grados de libertad es indistinguible a una normal, con lo que: $t_{1220;0.025} = 1.96$.

APARTADO C)

En este caso nos piden estimar $m_0 = E(Y_0|X_0 = 7)$, y sabemos que el intervalo de confianza para este parámetro está definido como:

$$IC_{0.95}(m_0) = \left[\hat{m}_0 \mp t_{n-2;\frac{\alpha}{2}} \cdot S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\right]$$

$$\hat{Y}_0 = \hat{m}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 1.4 + 0.61 \cdot 7 = 5.67$$
$$S_R = 1.016, \overline{x} = 6.60, S_{xx} = (n-1) \cdot S_x = 1221 \cdot 1.03^2$$
$$\sqrt{\frac{1}{1220} + \frac{(7-6.6)^2}{1221 \cdot 1.03}} = 0.31$$

S_x sabemos que es 1.03^2 porque S_x es la cuasivarianza y en el enunciado nos dan la cuasi-desviación típica.

El resultado final es:

$$IC = [5.67 \mp \underbrace{(1.96)(1.016)(0.031)}_{0.06}]$$

Ejercicio 3.2: *Dada una muestra de 10 observaciones, se ha ajustado un modelo de regresión simple por mínimos cuadrados, resultando:*

$$Y_i = 1 + 3x_i,\ R^2 = 0.9,\ S_R^2 = 2$$

Calcula un intervalo de confianza para la pendiente de la recta con un nivel de confianza 0.95. ¿Podemos rechazar, con un nivel de significación de 0.05, la hipótesis nula de que la variable x no influye linealmente en la variable Y?

Solución de clase:

Con los datos del ejercicio tendremos:

$$S_R^2 = 2 = \frac{SCE}{n-2} \implies SCE = 2 \cdot 8 = 16$$

y también:

$$R^2 = 0.9 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{16}{SCT} \implies SCT = 160$$

Para obtener el error típico de $\hat{\beta}_1$ necesitamos obtener $\sqrt{S_{xx}}$:

$$SCR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(\overline{Y} + \hat{\beta}_1(x_i - \overline{x}) - \overline{Y})^2 = \hat{\beta}_1^2 \cdot S_{xx}$$
$$\implies S_{xx} = \frac{SCR}{\hat{\beta}_1^2} = \frac{SCT - SCE}{9} = \frac{144}{9} = 16$$

De modo que ya podemos calcular $ET(\hat{\beta}_1) = \frac{S_R}{\sqrt{S_{xx}}} = \frac{\sqrt{2}}{4} \approx 0.35$, y por tanto nuestro intervalo de confianza para β_1 será:

$$IC_{0.95}(\beta_1) = \left[\hat{\beta}_1 \mp t_{8,0.025} \cdot ET(\hat{\beta}_1)\right] = [3 \mp 0.8152]$$

¿Podemos rechazar, con un nivel de significación de 0.05, la hipótesis nula de que la variable x no influye linealmente en la variable Y?

Para este contraste tendremos $H_0 : \beta_1 = 0$, y si nos construimos una tabla nos resultará más fácil llegar al estadístico F que necesitamos para hallar la región de rechazo:

Fuente	SC	gl	CM	F
Explicada	144	1	144	72
No explicada	16	8	2	
Total	160	9		

Sabemos que $R = \{F > F_{1,8;0.05}\}$, y puesto que $72 = F > F_{1,8;0.05}$ rechazamos H_0.

Hecho por Jorge. Se aceptan correcciones.

A la vista del modelo de regresión lineal presentado en el enunciado tendremos $\hat{\beta}_0 = 1$ y $\hat{\beta}_1 = 3$. Sabemos que un intervalo de confianza 0.95 para β_1 es:

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 \mp t_{n-2;\frac{\alpha}{2}} \frac{S_R}{\sqrt{S_{xx}}}\right]$$

Jorge: me imagino que con R se refiere a $S_{xx} = \frac{\sum_i (x_i - \overline{x})^2}{n}$, porque si no, no se me ocurre cómo calcularla sin saber $\overline{x}$ ni cada x_i.

Tenemos que $t_{8;0.025} = 3.83$, por lo que el intervalo de confianza queda:

$$IC_{0.95}(\beta_1) = \left[3 \mp 3.83 \cdot \frac{\sqrt{2}}{0.94}\right]$$

Veamos ahora si podemos decir que la variable x no influye linealmente en la variable Y ($H_0 : \beta_1 = 0$):

Sabemos que $\frac{\hat{\beta}_1 - \beta_1}{S_R/\sqrt{S_{xx}}} \equiv t_{n-2}$ sigue una t-student con n-2 grados de libertad, y bajo H_0 tendremos que $\frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} \equiv t_{n-2}$. Si queremos rechazar H_0 con nivel de significação $\alpha = 0.05$ la región de rechazo será:

$$R = \left\{\frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} > t_{n-2;\frac{\alpha}{2}}\right\} = \left\{\frac{3}{\sqrt{2}/0.94} > t_{8;0.025}\right\} = \{1.5 > 3.83\}$$

Por tanto **no caemos en la región de rechazo** que nos permitiría afirmar que x inluye linealmente en la variable Y.

Hecho por Dejuan. Se aceptan correcciones.

Lo primero es saber qué es R^2. En el ejercicio anterior, vemos que hay un "Adjusted R-squared". Gracias a nuestro conocimiento del inglés, R-squared es R^2, lo que nos conduce a pensar que ese R^2 es el "adjusted r-squared". La definicion dice

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{S_R(n-2)}{S_{xx}}$$

Entonces, despejamos S_{xx} de la ecuación:

$$0.9 = 1 - \frac{S_R(n-2)}{S_{xx}} = 1 - \frac{\sqrt{2}\cdot 8}{S_{xx}} \to 0.1 = \frac{16}{S_{xx}} \to S_{xx} = 160$$

Ahora ya podemos construir el intervalo de confianza:

$$IC_{1-\alpha}(\beta_1) = \left[\hat{\beta}_1 \mp t_{n-2;\frac{\alpha}{2}} \frac{S_R}{S_{xx}}\right] = \left[3 \mp 3.83 \cdot \frac{\sqrt{2}}{160}\right] = [3 \mp 0.034]$$

Veamos ahora si podemos decir que la variable x no influye linealmente en la variable Y ($H_0 : \beta_1 = 0$): Deberíamos poder rechazar (y por bastante), ya que si nuestra estimación es $\hat{\beta}_1 = 3$ y en realidad es 0... vaya mierda de estimación hemos hecho. Además, que $R^2 = 0.9$ valor cercano a 1 (valor máximo que puede tomar) también nos dice que el modelo construido es muy bueno.

Sabemos que $\frac{\hat{\beta}_1 - \beta_1}{S_R/\sqrt{S_{xx}}} \equiv t_{n-2}$ sigue una t-student con n-2 grados de libertad, y bajo H_0 tendremos que $\frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} \equiv t_{n-2}$. Si queremos rechazar H_0 con nivel de significación $\alpha = 0.05$ la región de rechazo será:

$$R = \left\{\frac{\hat{\beta}_1}{S_R/\sqrt{S_{xx}}} > t_{n-2;\frac{\alpha}{2}}\right\} = \left\{\frac{3}{\sqrt{2}/160} > t_{8;0.025}\right\} = \{339.41 > 3.83\} \implies$$

Ejercicio 3.3:

3. Supongamos que la muestra $(x_1, Y_1), \ldots, (x_n, Y_n)$ procede de un modelo de regresión lineal simple en el que se verifican las hipótesis habituales. Consideramos el siguiente estimador de la pendiente del modelo (se supone $x_1 \neq \overline{x}$):

$$\tilde{\beta}_1 = \frac{Y_1 - \overline{Y}}{x_1 - \overline{x}}$$

a) *¿Es $\tilde{\beta}_1$ un estimador insesgado?*

b) *Calcula la varianza de $\tilde{\beta}_1$.*

c) *Supongamos que la varianza de los errores del modelo, σ^2, es un parámetro conocido. Escribe la fórmula de un intervalo de confianza de nivel $1 - \alpha$ para β_1 cuyo centro sea el estimador $\tilde{\beta}_1$.*

Hecho por Jorge. Se aceptan correcciones.

Corregido en clase, aunque el apartado b se ha hecho de otra manera

APARTADO A)

Para este cálculo utilizamos:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i + \mathbb{E}(\varepsilon_i) = \beta_0 + \beta_1 x_i$$

Ya que $\varepsilon_i \equiv N(0, \sigma^2)$

Además, como las x son constantes: $\mathbb{E}(x_1 - \overline{x}) = x_1 - \overline{x}$.

Vamos a calcular el sesgo:

$$\mathbb{E}\left(\tilde{\beta}_1\right) = \frac{1}{x_1 - \overline{x}}(\mathbb{E}(Y_1) - \mathbb{E}\left(\overline{Y}\right)) = \frac{1}{x_1 - \overline{x}}(\beta_0 + \beta_1 x_1 - \mathbb{E}\left(\overline{Y}\right))$$

Vamos a ver el valor de $\mathbb{E}\left(\overline{Y}\right)$:

$$\mathbb{E}\left(\overline{Y}\right) = \frac{1}{n}\sum_{i=0}^{n}\mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=0}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \overline{x}$$

Por tanto al sustituir en la primera ecuación de este apartado obtenemos que $\mathbb{E}\left(\tilde{\beta}_1\right) = \beta_1$, y por tanto el estimador es insesgado.

APARTADO B)

$$\mathbb{V}\left(\tilde{\beta}_1\right) = \mathbb{V}\left(\frac{Y_1 - \overline{Y}}{x_1 - \overline{x}}\right) = \frac{1}{(x_1 - \overline{x})^2}\left[\mathbb{V}(Y_1) + \mathbb{V}\left(\overline{Y}\right) - 2\text{cov}\left(Y_1, \overline{Y}\right)\right]$$

Ya sabemos que en el modelo de regresión lineal $\mathbb{V}(Y_i) = \sigma^2,\ \forall i$, luego lo siguiente que haremos es calcular los otros dos términos del corchete por separado:

$$\mathbb{V}\left(\overline{Y}\right) = \mathbb{V}\left(\frac{\sum Y_i}{n}\right) \overset{Y_i \text{ independientes}}{=} \frac{1}{n^2}\sum \mathbb{V}(Y_i) = \frac{\sigma^2}{n}$$

Ahora miramos la covarianza:

$$\text{cov}\left(Y_1, \overline{Y}\right) = \text{cov}\left((1,0,0,\ldots,0)\overrightarrow{Y}, \frac{1}{n}(1,1,1,\ldots,1)\overrightarrow{Y}\right) = (1,\ldots,0)\cdot\sigma^2 I\cdot\frac{1}{n}\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \frac{\sigma^2}{n}$$

Y sustituyendo en la primera ecuación del apartado obtenemos que:

$$\mathbb{V}\left(\tilde{\beta}_1\right) = \frac{\sigma^2}{(x_1 - \overline{x})^2}\left(1 - \frac{1}{n}\right)$$

APARTADO C)

Puesto que podemos expresar $\tilde{\beta}_1$ como:

$$\tilde{\beta}_1 = \frac{1}{x_1 - \overline{x}} \left((1, \dots, 0) - \frac{1}{n}(1, \dots, 1) \right) \cdot \overrightarrow{Y}$$

Donde:

$$\overrightarrow{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

es un vector de normales Y_i independientes. Así que podemos decir que $\tilde{\beta}_1$ es una combinación lineal de normales, y por tanto seguirá una distribución normal:

$$\tilde{\beta}_1 \equiv N \left(\beta_1, \underbrace{\frac{\sigma^2}{(x_1 - \overline{x})^2} \left(1 - \frac{1}{n} \right)}_{v} \right)$$

Por tanto $\frac{\tilde{\beta}_1 - \beta_1}{\sqrt{v}} \equiv N(0,1)$, y podemos definir el intervalo de confianza:

$$IC_{1-\alpha}(\beta_1) = \left[\tilde{\beta}_1 \mp Z_{\frac{\alpha}{2}} \cdot \sqrt{v} \right]$$

Si te preguntas porqué es $\mathcal{Z}$ y no $\mathcal{T}$, revisa la construcción del intervalo de confianza para β_1 (en 1.1.6)

Ejercicio 3.4: *Se considera el siguiente modelo de regresión simple a través del origen:*

$$Y_i = \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \equiv N(0, \sigma^2) \text{ independientes},\ i = 1, ..., n.$$

a) *Calcula el estimador de mínimos cuadrados de β_1 y deduce su distribución.*

b) *Sean $e_i,\ i = 1, ..., n$ los residuos del modelo. Comprueba si se cumplen o no las siguientes propiedades: $\sum_{i=1}^n e_i = 0$ y $\sum_{i=1}^n e_i x_i = 0$.*

c) *Si la varianza de los errores σ^2 es conocida, deduce la fórmula de un intervalo de confianza de nivel $1 - \alpha$ para el parámetro β_1.*

APARTADO A)

Entonces, $\Phi(\beta) = \sum (y_i - \beta x_i)^2$. Derivando e igualando a 0 se llega a $\sum (y_i - \hat{\beta}_1 x_i) x_i = 0$, y obtenemos el estimador despejando $\hat{\beta}_1$.

Otra manera de hacerlo es utilizando lo que hemos visto en regresión múltiple para modelos lineales, definiendo la matriz de diseño X como $\hat{\beta}_1 = (X'X)^{-1}X'Y$

En ambos casos se llega a:

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Vamos a calcular su esperanza y su varianza para la distribución:

$$\mathbb{E}\left(\hat{\beta}\right) = \beta$$

$$\mathbb{V}\left(\hat{\beta}\right) = \frac{\sum x_i^2 \sigma^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\sum x_i^2}$$

APARTADO B)

Como no hay término independiente, los residuos no suman 0. Esto tiene varios razonamientos intuitivos.

Si en la matriz de diseño no hay una columna que sea todo 1's, (porque no haya término independiente) entonces el vector de residuos no es ortogonal a V.

Sin embargo en este caso se cumple que $\sum e_i x_i = 0$ pues al minimizar Φ se ha obtenido que $\sum \underbrace{(y_i - \hat{\beta}_1 x_i)}_{e_i} x_i = 0$

APARTADO C)

$$IC_{1-\alpha}(\hat{\beta}_1) = \left[\hat{\beta}_1 \mp Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{\sum x_i^2}}\right]$$

Si en el enunciado no nos dijeran que conocemos σ, tendríamos que cambiar σ por S_R que es un dato que sí tenemos. Entonces, construiríamos:

$$IC_{1-\alpha}(\hat{\beta}_1) = \left[\hat{\beta}_1 \mp t_{n-2;\frac{\alpha}{2}} \cdot \frac{S_R}{\sqrt{\sum x_i^2}}\right]$$

Ejercicio 3.5: *En el modelo del problema anterior supongamos que $x_i > 0$ y que $\mathbb{V}(\varepsilon_i) = \sigma^2 x_i^2$, es decir, no se cumple la hipótesis de homocedasticidad. Calcula en este caso la esperanza y la varianza del estimador de mínimos cuadrados $\hat{\beta}_1$. Consideremos ahora el estimador alternativo $\tilde{\beta}$ que se obtiene al minimizar la expresión $\sum_{i=1}^n w_i (y_i - \beta_1 x_i)^2$, donde $w_i = 1/x_i^2$. Calcula una fórmula explícita para $\tilde{\beta}$ y, a partir de ella, deduce su esperanza y su varianza. Compara los estimadores $\hat{\beta}_1$ y $\tilde{\beta}$. ¿Cuál es mejor? (A $\tilde{\beta}_1$ se le llama estimador de mínimos cuadrados ponderados).*

a) *¿insesgado?*

APARTADO A)

Es razonable que sea insesgado, ya que en media sí puede tener sentido. El problema será la varianza... vamos a calcular la distribución del estimador de mínimos cuadrados:

Como hemos calculado en el ejercicio anterior:

$$\mathbb{E}(\beta_1) = \beta_1$$

$$\mathbb{V}(\beta_1) = \mathbb{V}\left(\frac{\sum x_i y_i}{\sum x_i^2}\right) = \ldots = \sigma^2 \frac{\sum x_i^4}{\left(\sum x_i^2\right)^2}$$

Vamos a pensar... ¿De qué puntos nos podemos fiar más? ¿De los pequeños o de los grandes? Al ser heterocedástico, donde menor varianza hay es en los x_i cercanos al origen, con lo que deberíamos fiarnos más de ellos. Esta "confianza" la implementamos con una ponderación, obteniendo el **Mínimos cuadrados ponderados**

Mínimos cuadrados ponderados

Los cálculos se dejan para el lector, aunque el resultado será:

- Ambos son insesgados.
- En términos de varianza, es mejor el ponderado.
- ¿Cuál es el problema de ponderar? Que no sabemos con exactitud que $\varepsilon_i \sim N(0, \sigma^2 x_i^2)$. ¿Y si fuera $\varepsilon_i \sim N(0, \sigma^2 x_i^4)$? Entonces no podríamos aplicar los pesos calculados y es muy problemático en ese sentido.

"Cálculos para el lector"

Hecho por Jorge. Se aceptan correcciones.

Tenemos:

$$\Phi(\beta_1) = \sum_{i=1}^{n} \frac{1}{x_i^2}(y_i - \beta_i x_i)^2$$

$$\frac{\partial \Phi}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(\frac{y_i}{x_i} - \beta_1\right) = 0$$

De modo que el $\hat{\beta}_1$ que minimiza $\Phi(\beta_1)$ será:

$$\tilde{\beta}_1 = \frac{1}{n}\sum \frac{y_i}{x_i}$$

Veamos que el estimador $\tilde{\beta}_1$ es insesgado:

$$\mathbb{E}(\tilde{\beta}_1) = \frac{1}{n}\sum \frac{1}{x_i}\mathbb{E}(y_i) = \frac{1}{n}\sum \frac{\beta_1 x_i}{x_i} = \beta_1$$

Ahora calculamos su varianza:

$$\mathbb{V}\left(\tilde{\beta}_1\right) \underbrace{=}_{y_i \text{ indeps.}} \frac{1}{n^2}\sum \frac{1}{x_i^2}\mathbb{V}(y_i) = \frac{\sigma^2}{n}$$

Ejercicio 3.6: *Supongamos que cierta variable respuesta Y depende linealmente de dos variables regresoras x_1 y x_2, de manera que se verifica el modelo:*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, i = 1, \ldots, n,$$

donde los errores εi verifican las hipótesis habituales. Se ajusta por mínimos cuadrados el modelo $Y_i = \beta_0 + \beta_1 x_{i1}$, sin tener en cuenta la segunda variable regresora. Demuestra que el estimador β_1 es, en general, sesgado y determina bajo qué condiciones se anula el sesgo.

Sabemos que:

$$\hat{\beta}_1 = \frac{S_{x_1 y}}{S_{x_1 x_1}} = \frac{\sum (x_{i1} - \overline{x}_1) y_i}{\sum (x_{i1} - \overline{x}_1)^2}$$

Por tanto el valor esperado del estimador será:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \frac{1}{S_{x_1 x_1}} \sum (x_{i1} - \overline{x}_1) \cdot (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) = \\
&= \frac{1}{S_{x_1 x_1}} \left[\beta_0 \cdot 0 + \beta_1 \sum (x_{i1} - \overline{x}_1)^2 + \beta_2 \sum (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) \right] \\
&= \beta_1 + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{S_{x_1 x_1}}
\end{aligned}$$

De modo que el estimador será insesgado cuando x_1 y x_2 sean independientes, ya que así se tendrá que $\operatorname{cov}(x_1, x_2) = 0$.

Ejercicio 3.7: *En el Ayuntamiento de Madrid se estudió hace unos años la conveniencia de instalar mamparas de protección acústica en una zona de la M-30. Un técnico del Ayuntamiento piensa que si el ruido afecta mucho a los habitantes de la zona esto debe reflejarse en los precios de las viviendas. Su idea es que el precio de una casa en esa zona (y) depende del número de metros cuadrados (x_1), del número de habitaciones (x_2) y de la contaminación acústica, medida en decibelios, (x_3). Para una muestra de 20 casas vendidas en los últimos tres meses, se estima el siguiente modelo:*

$$\hat{y}_i = \underset{}{5970} + \underset{(2,55)}{22,35} x_{i1} + \underset{(1820)}{2701,1} x_{i2} - \underset{(15,4)}{67,6730} x_{i3}$$

$$R^2 = 0,9843$$

donde las desviaciones típicas (estimadas) de los estimadores de los coeficientes aparecen entre paréntesis.

***a)** Calcula el efecto que tendría sobre el precio un descenso de 10 decibelios, si el resto de variables en el modelo permanecieran constantes.*

***b)** Contrasta con $\alpha = 0,05$ la hipótesis nula de que el número de habitaciones no influye en el precio.*

***c)** A nivel $\alpha = 0,05$, ¿puede afirmarse que la vivienda se encarece cuando disminuye la contaminación acústica?*

***d)** Contrasta con $\alpha = 0,05$ la hipótesis nula de que las tres variables no influyen conjuntamente en el precio.*

***e)** Estima el precio medio de las casas (no incluidas en la muestra) que tienen 100 metros cuadrados, dos habitaciones y una contaminación acústica de 40 decibelios.*

APARTADO A)

Fijando las variables y haciendo $x_{i3} \to x_{i3} - 10$ se ve que $\hat{y}_i$ se incrementaría en 676.73.

APARTADO B)

$H_0 : \beta_2 = 0$ y sabemos que $\hat{\beta}_2 \sim N(\beta_2, \sigma^2 q_{22})$. Por tanto nos servimos del contraste:

$$\frac{\hat{\beta}_i}{\text{desv-estim}(\hat{\beta}_i)} \sim t_{n-k-1=16}$$

$$t = \frac{2701.1}{1820} = 1.4841,\ t_{16;0.025} = 2.12$$

Como $t < t_{16;0.025}$ aceptamos H_0 (el número de habitaciones no influye en el precio).

APARTADO C)

Para este contraste nos basamos en el visto en teoría 1.1.7 (aunque ahora la hipótesis alternativa es lo contrario) y establecemos como hipótesis alternativa lo que queremos afirmar $H_1 : \beta_3 < 0$, y como hipótesis nula $H_0 : \beta_3 \geq 0$. De este modo lo que queremos ver es que nuestro estadístico cae en la región de rechazo, es decir, queremos que, siendo negativo, sea más pequeño que $-t_{16;0.05} = -1.74$:

$$t = \frac{\hat{\beta}_3}{\text{desv-estim}(\hat{\beta}_3)} = \frac{-67.673}{15.4} = -4.39$$

Como $t < -t_{16;0.06}$ rechazamos H_0, y por tanto aceptamos que los precios suben cuando se disminuyen los decibelios (H_1).

APARTADO D)

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, es decir, vamos a llevar a cabo el contraste conocido como "contraste de la regresión":

$F = \frac{SCR/R}{SCE/(n-k-1)} = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = 334.369 > F_{3,16;0.005} = 3.23$

Por tanto rechazamos H_0.

APARTADO E)

Nos piden $\hat{m}_0 = \hat{\beta} X_0 = (5970, 21.3, 2701.1, -67.67) \cdot (1, 100, 2, 40)' = 10900.28$, tal y como se explica en 1.4.

Ejercicio 3.8: *Se desea ajustar el modelo $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, donde los errores ε_i verifican las hipótesis habituales en el modelo de regresión múltiple. Los datos disponibles de las variables regresoras y la variable respuesta se encuentran en la matriz de diseño X y vector Y siguientes:*

$$X = \begin{pmatrix} 1 & 1 & -2 \\ 1 & -1 & 2 \\ 1 & 2 & 1 \\ 1 & -2 & -1 \end{pmatrix}, Y = \begin{pmatrix} 2 \\ 2 \\ 4.5 \\ -4.5 \end{pmatrix},$$

a) *Calcula los estimadores de mínimos cuadrados de $\beta_0, \beta_1 y \beta_2$.*

b) *Sabiendo que la varianza residual es $S_R^2 = 0.25$, contrasta la hipótesis nula $H_0 : \beta_1 = 0$.*

APARTADO A)

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1}X' \cdot Y = \begin{pmatrix} 1 \\ 1.8 \\ 0.9 \end{pmatrix}$$

APARTADO B)

Para este contraste nos apoyamos en que $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 q_{11})$, donde q_{11} es la entrada de la matriz $(X'X)^{-1}$ asociada a $\hat{\beta}_1$. Esto es así porque sabemos que $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. En este tipo de contraste estimamos σ^2 con S_R^2 y obtenemos que bajo $H_0 : \beta_1 = 0$:

$$\frac{\left|\hat{\beta}_1\right|}{S_R\sqrt{q_{11}}} \sim t_{n-k-1=1}$$

Echando cuentas se obtiene que:

$$(X'X)^{-1} = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/10 & 0 \\ 0 & 0 & 1/10 \end{pmatrix},$$

Y por tanto $q_{11} = 1/10$, lo cual permite hacer el contraste:

$$t = \frac{\left|\hat{\beta}_1\right|}{S_R\sqrt{q_{11}}} = 11.38$$

Que en un nivel de significación α habitual implicaría un rechazo de H_0 (es decir $t > t_{1;\frac{\alpha}{2}}$).

Ejercicio 3.9: *Se considera el siguiente modelo de regresión lineal múltiple:*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,\ \varepsilon_i \equiv N(0, \sigma^2) \qquad \text{(A.3.1)}$$

Se dispone de $n = 20$ observaciones con las que se ajustan todos los posibles submodelos del modelo A.3.1, obteniéndose para cada uno de ellos las siguientes sumas de cuadrados de los errores (todos los submodelos incluyen un término independiente).

Variables incluidas en el modelo	*SCE*	*Variables incluidas en el modelo*	*SCE*
Sólo término independiente	*42644.00*	x_1 *y* x_2	*7713.13*
x_1	*8352.28*	x_1 *y* x_3	*762.55*
x_2	*36253.69*	x_2 *y* x_3	*32700.17*
x_3	*36606.19*	x_1, x_2 *y* x_3	*761.41*

(***Ejemplo en negrita**: Para el modelo ajustado $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$, la suma de cuadrados de los errores es 32700.17).*

a) *Calcula la tabla de análisis de la varianza para el modelo A.3.1 y contrasta a nivel* α *= 0,05 la hipótesis nula* $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

b) *En el modelo A.3.1, contrasta a nivel* $\alpha = 0.05$ *las dos hipótesis nulas siguientes:*

- $H_0 : \beta_2 = 0$
- $H_0 : \beta_1 = \beta_3 = 0$

c) *Calcula el coeficiente de correlación entre la variable respuesta y la primera variable regresora sabiendo que es positivo.*

OJO : en clase dijo que este era uno de los problemas difíciles de un control

APARTADO A)

Bajo $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ tendremos que $Y_i = \beta_0 + \varepsilon_i$ y que $\hat{\beta}_0 = \overline{Y}$, y por tanto:

$$SCE_0 = \sum (Y_i - \hat{Y}_i)^2 \underbrace{=}_{H_0} \sum (Y_i - \overline{Y})^2 = SCT$$

En este caso tenemos que llevar a cabo el cálculo del estadístico del contraste de la regresión (véase 2.3) $F = \frac{SCR/k}{SCE/(n-k-1)}$. Como sabemos que $SCT = SCE + SCR \implies SCR = SCE_0 - SCE = 42644.00 - 761.41 = 41882.59$ podemos obtener la tabla con la que conseguimos el estadístico:

Fuente	SC	gl	CM	F
Explicada	$SCR = 41882.59$	$k = 3$	13960.86	293.37
No explicada	$SCE = 761.41$	$n - k - 1 = 16$	$47.59 = S_R^2$	
Total	42644	19		

Sabemos que la región de rechazo será: $R = \{F > F_{3,16;0.05} = 3.24\}$, y por tanto **rechazamos** H_0.

APARTADO B)

- $H_0 : \beta_2 = 0$. En este caso contrastamos el incremento de variabilidad relativa entre el modelo en el que solo tenemos en cuenta x_1, x_3, frente al modelo completo en el que tenemos en cuenta x_1, x_2, x_3:

$$F = \frac{\frac{SCE_0 - SCE}{p=1}}{\frac{SCE}{n-k-1}} = \frac{SCE_0 - SCE}{S_R^2} = \frac{762.55 - 761.41}{47.59} \approx 0.024$$

  En este caso la región de rechazo es $R = \{F > F_{1,16;0.05} = 4.49\}$, y por tanto no rechazamos la hipótesis nula H_0.

- $H_0 = \beta_1 = \beta_3 = 0$, aplicando el mismo criterio que en caso anterior obtenemos:

$$F = \frac{\frac{SCE_0 - SCE}{2}}{S_R^2} = \frac{\frac{36253.69 - 761.41}{2}}{47.59} = 372.9$$

Puesto que $F_{2,16;0.05} = 3.63$, rechazamos esta hipótesis nula.

APARTADO C)

Correlación entre Y y x_1:

$$r^2 = R^2 = 1 - \frac{SCE}{SCT} = 1 - \frac{8352.28}{42644} = 0.8041$$

De modo que tendremos $r = \pm\sqrt{0.8041}$, y con la ayuda del enunciado podemos decir que $r = +\sqrt{0.8041} = 0.8967$

Ejercicio 3.10: *A partir de una muestra de $n = 20$ observaciones se ha ajustado el modelo de regresión lineal simple $Y_i = 0 + 1x_i + \varepsilon_i$ con los siguientes resultados:*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29016    1.66161   0.175    AAA
x            1.01450    0.03246  31.252   <2e-16
Residual standard error: 0.1717 on 18 degrees of freedom
Multiple R-squared: 0.9819,     Adjusted R-squared: 0.9809
F-statistic: BBB on 1 and 18 DF,  p-value: < 2.2e-16
> vcov(reg)
            (Intercept)     x
(Intercept)  2.761      -0.054
x           -0.054       0.001
```

---

a) *Determina si el p-valor AAA es mayor o menor que 0.1. Escribe la hipótesis nula a la que corresponde este p-valor y determina si esta hipótesis se rechaza o no a nivel $\alpha = 0.1$.*

b) *Contrasta la hipótesis nula $H_0 : \beta_0 + \beta_1 = 2$ a nivel $\alpha = 0.05$.*

c) *Calcula el valor BBB que se ha omitido en los resultados anteriores.*

APARTADO A)

Si queremos ver qué t_{18} se corresponde con el nivel de significación $\alpha = 0.1$ buscamos en las tablas $t_{18,0.05} = 1.734$, que vemos que es claramente mayor que el $t = 0.175$ obtenido en la salida de R. Por tanto $P\{|t| > 0.175\} > 0.1$, lo que quiere decir que el p-valor AAA es mayor que 0.1.

La hipótesis nula asociada al p-valor AAA es $H_0 : \beta_0 = 0$ y puesto que su p-valor es menor que $t_{18,\alpha/2=0.1/2=0.05}$, no rechazamos H_0.

APARTADO B)

$\beta_0 + \beta_1 = a'\beta = (1, 1) \cdot \beta$. Multiplicando por a tenemos que:

$$a'\hat{\beta} \sim N\left(a'\beta, \sigma^2 a'(X'X)^{-1}a\right) \implies \frac{a'\hat{\beta} - a'\beta}{\sigma^2\sqrt{a'(X'X)^{-1}a}} \sim N(0, 1)$$

y puesto que en este apartado $H_0 : a'\beta = 2$, tras aproximar σ^2 por S_R^2 queda que:

$$t = \frac{\left|a'\hat{\beta} - 2\right|}{S_R\sqrt{a'(X'X)^{-1}a}} \sim t_{n-k-1=18}$$

La salida *vcov(reg)* que figura en la salida de R es la estimación de la matriz de covarianzas de $\hat{\beta}$ (la matriz de la que hablamos es $\sigma^2(X'X)^{-1}$), que no es más que $S_R^2(X'X)^{-1}$. Por tanto:

$$S_R^2 a'(X'X)^{-1}a = a'S_R^2(X'X)^{-1}a = 2.65$$

$$t = \frac{\left|a'\hat{\beta} - 2\right|}{S_R\sqrt{a'(X'X)^{-1}a}} = \frac{|-0.6955|}{\sqrt{2.65}} = 0.4287 < t_{18;0.025}$$

De modo que no rechazamos $H_0 : \beta_0 + \beta_1 = 2$.

APARTADO C)

La última línea corresponde con el contraste de la regresión, es decir, $H_0 : \beta_1 = 0$. El estadístico para este contraste se obtiene como:

$$BBB = F = \frac{SCE_0 - SCE/k}{SCE/(n-k-1)} = \frac{SCR}{SCE/18}$$

Pero acabamos antes recordando la identidad $F = \frac{R^2}{1-R^2}\frac{n-k-1}{k}$ (obviamente equivalente a lo anterior) para obtener que:

$$BBB = F = \frac{0.9819}{1-0.9819} \cdot 18 = 976.47$$

Ejercicio 3.11: *Se desea estudiar la esperanza de vida Y en una serie de países como función de la tasa de natalidad nat, la tasa de mortalidad infantil mortinf y el logaritmo del producto nacional bruto lpnb. Para ajustar el modelo*

$$Y_i = \beta_0 + \beta_1 \cdot nat_i + \beta_2 \cdot mortinf_i + \beta_3 \cdot lpnb_i + \varepsilon_i$$

donde los errores ε_i son v.a.i.i.d. $N(0, \sigma^2)$. Se ha utilizado R con los siguientes resultados:

```
> reg = lm(Y~nat+mortinf+lpnb)
> summary(reg)
Call:
lm(formula = Y ~ nat + mortinf + lpnb)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.24045    2.90253  23.855  < 2e-16
nat         -0.17572    0.04244  -4.140    8e-05
mortinf     -0.14086    0.01370 -10.284  < 2e-16
lpnb         0.98901    0.29404   3.363  0.00115
---
Residual standard error: 2.788 on 87 degrees of freedom
Multiple R-Squared: 0.9303,     Adjusted R-squared: 0.9279
F-statistic: 386.9 on 3 and 87 DF,  p-value: < 2.2e-16

> anova(reg)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
nat        1 7602.7  7602.7 977.798 < 2.2e-16
mortinf    1 1334.2  1334.2 171.599 < 2.2e-16
lpnb       1   88.0    88.0  11.313  0.001146
Residuals 87  676.5     7.8
---
```

a) *¿De cuántos países consta la muestra utilizada?*

b) *¿Cuál es la suma de cuadrados de la regresión (SCR) que se utiliza para medir la variabilidad explicada por las tres variables regresoras?*

c) *¿Cuánto vale la cuasivarianza muestral de la variable respuesta* $\sum(Y_i-\overline{Y})^2/(n-1)$*?*

d) *Contrasta a nivel* α *= 0,05 la hipótesis nula* $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

e) *Determina cuál es la hipótesis nula y la alternativa correspondiente a cada uno de los tres estadísticos F que aparecen en la tabla de análisis de la varianza anterior.*

Hecho por Dejuan. Se aceptan correcciones.

APARTADO A)

Sabemos $n-k-1=87$ y $k=3$, con lo que $n=91$

APARTADO B)

SCE=676.5 y $R^2=0.93$. Utilizando $R^2=1-\frac{SCE}{SCT}$ despejamos $SCT=9705.88$ y con este, obtenemos $SCR=9029.38$

APARTADO C)

$$\frac{SCT}{n-1}=99.224$$

APARTADO D)

$$R=\left\{385.9>F_{3,18}\right\}$$

Podemos observar que el p-valor del contraste de la regresión (en que nos piden) es cercano a 0, con lo que rechazamos H_0.

APARTADO E)

El primero corresponde al contraste de la regresión, el segundo corresponde al cpntraste $\beta_2, \beta_3 =$ y el tercero corresponde a $\beta_3 = 0$.

Ejercicio 3.12: *Considera el modelo de regresión múltiple $Y = X\beta + \varepsilon$, donde ε verifica las hipótesis habituales.*

***a)** Define el vector de valores ajustados $\hat{Y} = (\hat{Y_1}, ..., \hat{Y_n})$*

***b)** En general, ¿son las variables $\hat{Y_1}, ..., \hat{Y_n}$ independientes? ¿Son idénticamente distribuidas?*

***c)** Calcula el valor de $\sum_{i=1}^{n} \mathbb{V}\left(\hat{Y_i}\right)$ si el modelo incluye un término independiente y 3 variables regresoras.*

APARTADO A)

$$\mathbb{E}(\hat{Y}) = X\hat{\beta} = HY$$
$$\hat{Y} \equiv N(X\beta, \sigma^2 H)$$

Sabemos que esa es la varianza porque $Y \equiv N_n(X\beta, \sigma^2 I_n)$ y aplicamos $\hat{Y} = AY \rightarrow \mathbb{V}\left(\hat{Y}\right) = A\Sigma A'$

APARTADO B)

No son independientes en general porque H no es siempre diagonal. Tampoco son idénticamente distribuidas porque no tienen la misma varianza ni la misma media:

$$\mathbb{V}\left(\hat{Y_i}\right) = \sigma^2 h_{ii}$$

Donde h_{ii} es el potencial de la i-ésima observación.

APARTADO C)

Lo que nos piden es la traza de H. Como H es idempotente, tenemos $\sigma^2 \operatorname{traza}(H) = \sigma^2 \operatorname{Rg}(H) = \sigma^2(k+1) = 4\sigma^2$

Sabemos que $\operatorname{Rg}(H) = k + 1$ por hipótesis (tenemos k variables más el término independiente). En este caso $k = 3$.

Ejercicio 3.13:

Con el fin de evaluar el trabajo de los directores de los 30 departamentos de una gran empresa, se llevó a cabo una encuesta a los empleados a su cargo en la que se les pidió que valoraran varias afirmaciones con una nota de 1 (máximo acuerdo) a 5 (máximo desacuerdo). Algunas de las variables eran: Y , el trabajo del director es en general satisfactorio; x 1

, el director gestiona correctamente las quejas de los empleados; x 2 , el director trata equitativamente a los empleados; x 3 , la asignación del trabajo es tal que los empleados pueden aprender cosas nuevas con frecuencia. El vector (Y i , x i1 , x i2 , x i3) contiene la suma de puntos de las respuestas en el departamento i, donde i = 1, . . . , 30. Con estos datos se ajustó con R el modelo:

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2583     7.3183   1.538   0.1360
x1            0.6824     0.1288   5.296 1.54e-05
x2           -0.1033     0.1293  -0.799   0.4318
x3            0.2380     0.1394   1.707   0.0997
---
Residual standard error: 6.863 on 26 degrees of freedom
Multiple R-squared: 0.715,      Adjusted R-squared: 0.6821
F-statistic: AAA on BBB and CCC DF,  p-value: 2.936e-07
```

a) *Calcula un intervalo de confianza de nivel 0.95 para el parámetro β_3 . Contrasta la hipótesis $H_0 : \beta 3 \leq 0$.*

b) *Determina el valor de AAA, BBB y CCC en la última línea de la salida anterior. ¿A qué hipótesis nula corresponde el p-valor que aparece en esta última línea?*

APARTADO A)

$$IC(\beta_3) = \left[0.238 \mp 0.1394 \cdot t_{28;0.025}\right] = [0.238 \mp 0.2855]$$

Y la región de rechazo correspondiente es:

$$R = \left\{ \frac{\beta_3}{e.t.(\beta_3)} > t_{n-2;\alpha} \right\} = \{1.707 > 1.701\}$$

Entonces rechazamos la hipótesis H_0.

Observación: Es curioso que rechazamos la hipótesis de que β_3 pueda ser negativo, pero uno de los extremos del intervalo de confianza es negativo.

APARTADO B)

Corresponde al contraste de la regresión que es $\beta_1 = \beta_2 = \beta_3 = 0$

Tenemos BBB=3, CCC=26 con lo que $AAA = \frac{R^2}{1-R^2} \cdot \frac{26}{3} = 21.74$

Ejercicio 3.14: *Tres vehículos se encuentran situados en los puntos $0 < \beta_1 < \beta_2 < \beta_3$ de una carretera recta. Para estimar la posición de los vehículos se toman las siguientes medidas (todas ellas sujetas a errores aleatorios de medición independientes con distribución normal de media 0 y varianza σ^2):*

- *Desde el punto 0 medimos las distancias a los tres vehículos dando $Y_1, Y_2 e Y_3$*

- *Nos trasladamos al primer vehículo y medimos las distancias a los otros dos, dando dos nuevas medidas Y_4, Y_5 .*
- *Nos trasladamos al segundo vehículo y medimos la distancia al tercero, dando una medida adicional Y_6.*

APARTADO A)

Expresa el problema de estimación como un modelo de regresión múltiple indicando clara- mente cuál es la matriz de diseño.

APARTADO B)

Calcula la distribución del estimador de mínimos cuadrados del vector de posiciones $(\beta_1, \beta_2, \beta_3)$.

APARTADO C)

Se desea calcular un intervalo de confianza de nivel 95 % para la posición del primer vehículo β_1 a partir de 6 medidas (obtenidas de acuerdo con el método descrito anteriormente) para las que la varianza residual resultó ser $S_R^2 = 2$. ¿Cuál es el margen de error del intervalo?

a)

$$\begin{aligned} Y_1 &= \beta_1 + \varepsilon_1 \\ Y_2 &= \beta_2 + \varepsilon_2 \\ Y_3 &= \beta_3 + \varepsilon_3 \\ Y_4 &= \beta_2 - \beta_1 + \varepsilon_4 \\ Y_5 &= \beta_2 - \beta_1 + \varepsilon_5 \\ Y_6 &= \beta_3 - \beta_2 + \varepsilon_6 \end{aligned}$$

Vamos a construir la matriz de diseño. Será de la forma:

$$Y = \left(X\right) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \varepsilon$$

De esta manera:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

Se ha dejado caer en clase un posible ejercicio de examen: ¿Cuál es la matriz de diseño óptima para estimar los β_i?

b) Con esta matriz de diseño, podemos calcular:

$$\hat{\beta} = N_3\left(\beta, \sigma^2 \underbrace{\begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix}}_{(X'X)^{-1}}\right)$$

c)

$$IC_{0.95}(\beta_1) \equiv \left[\hat{\beta}_1 \mp t_{6-3;0.025} S_R \sqrt{q_{11}}\right]$$

$$IC_{0.95}(\beta_1) \equiv \left[\hat{\beta}_1 \mp t_{6-3;0.025} \sqrt{2} \sqrt{\frac{1}{2}}\right]$$

Con lo que el margen de error es $t_{6-3;0.025}$

Ejercicio 3.15: *Sean Y_1, Y_2 e Y_3 tres variables aleatorias independientes con distribución normal y varianza 2. Supongamos que μ es la media de Y_1, λ es la media de Y_2 y $\lambda + \mu$ es la media de Y_3, donde, $\lambda, \mu \in \mathbb{R}$.*

***a)** Demuestra que el vector $Y = (Y_1, Y_2, Y_3)'$ verifica el modelo de regresión múltiple $Y = X\beta + \varepsilon$. Para ello, determina la matriz de diseño X, el vector de parámetros y la distribución de las variables de error ε.*

***b)** Calcula los estimadores de máxima verosimilitud (equivalentemente, de mínimos cuadrados) de λ, μ.*

***c)** Calcula la distribución del vector $(\hat{\lambda}, \hat{\mu})'$, formado por los estimadores calculados en el apartado anterior.*

APARTADO A)

$$Y = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \lambda \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

APARTADO B)

Tenemos una fórmula para calcularlo.

$$\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix} = (X'X)^{-1} X'Y$$

En caso de no sabernos la fórmula, podemos recurrir al método largo y tradicional:

$$\varphi(\lambda,\mu) = (Y_1-\mu)^2 + (Y_2-\lambda)^2 + (Y_3-(\lambda+\mu))^2$$

Y resolvemos el sistema:

$$\left.\begin{array}{ll} \frac{\partial\varphi}{\partial\lambda} & = 0 \\ \frac{\partial\varphi}{\partial\mu} & = 0 \end{array}\right\}$$

De esta manera deberíamos llegar a la misma solución.

$$\begin{pmatrix}\hat{\lambda}\\ \hat{\mu}\end{pmatrix} = \begin{pmatrix}\frac{2Y_2+Y_3-Y_1}{3}\\ \frac{2Y_1+Y_3-Y_2}{3}\end{pmatrix}$$

Podríamos comprobar si son insesgado o no.

APARTADO C)

Sabemos que la distribución del estimador es:

$$\hat{\beta} = \begin{pmatrix}\hat{\lambda}\\ \hat{\mu}\end{pmatrix} \equiv N\left(\beta, \sigma^2(X'X)^{-1}\right) = N\left(\begin{pmatrix}\lambda\\ \mu\end{pmatrix}, \frac{\sigma^2}{3}\begin{pmatrix}2 & -1\\ -1 & 2\end{pmatrix}\right)$$

Ejercicio 3.16:

La siguiente tabla contiene información sobre los resultados de un examen en cuatro grupos de una misma asignatura:

	Alumnos	*Media*	*Cuasi-varianza*
Grupo 1	*104*	*4.99*	*4.19*
Grupo 2	*102*	*4.63*	*5.75*
Grupo 3	*69*	*4.53*	*5.15*
Grupo 4	*80*	*4.79*	*5.35*

Se supone que se satisfacen las hipótesis del modelo unifactorial. Escribe la tabla de análisis de la varianza y contrasta la hipótesis de que las notas medias son iguales en los cuatro grupos, con un nivel de significación $\alpha = 0,05$.

$$Y_{1\cdot} = 4.99; S_1 = 4.19$$

Vamos a construir la tabla ANOVA. Para ello:

$$Y_{\cdot\cdot} \neq \frac{\sum Y_{i\cdot}}{4}$$

Ya que el número de alumnos es distinto en cada grupo. La media total sería:

$$Y_{\cdot\cdot} = \frac{\sum n_i Y_{i\cdot}}{\sum n_i}$$

Ahora podemos calcular $SCR = \sum_{i=1}^{4} n_i(Y_{i\cdot} - Y_{\cdot\cdot})^2 = ... = 10.93$

$$SCE = \sum_i \sum_j (Y_{ij} - \overline{Y_{i\cdot}})^2$$

Fuente	SC	gl	CM	F
SCR	10.93	$4-1$	$\frac{10.93}{3} = 3.64$	$\frac{3.64}{5.09} = 0.72$
SCE	1785.17	$n-k = 351$	$\frac{1785.17}{351} = 5.09$	

Ahora buscamos $F_{3,351;0.05} = 2.60 > 0.72$, por lo que no hemos encontrado diferencias significativas de que el grupo influya en la nota. Aceptamos H_0.

Ejercicio 3.17: *Una fabricante de botas para el agua está estudiando tres posibles colores para su nuevo modelo de bota super resistente. Las opciones que se proponen son verde, amarillo y azul. Para analizar si el color tiene algún efecto sobre el número de ventas, se eligen 16 tiendas de tamaño similar. Se envían las botas de color verde a 6 tiendas elegidas al azar, las amarillos a 5 tiendas y las azules a las 5 restantes. Después de varios días se comprueba el número de botas vendidas en cada tienda, obteniéndose los siguientes resultados:*

Verdes	*Amarillas*	*Azules*
43	*52*	*61*
52	*37*	*29*
59	*38*	*38*
76	*64*	*53*
61	*74*	*79*
81		

Es igual que el anterior. Se deja para otro.

A.4. Hoja 4

Ejercicio 4.1:

a) *Estima a partir de estos datos, la función lineal discriminante de Fisher.*

b) *Clasifica la observación* $xx = (2,7)'$ *utilizando la regla obtenida en el apartado anterior.*

APARTADO A)

Vamos a estimar las medias de cada población:

$$\hat{\mu_0} = \overline{x_0} = (3,6)'$$

$$\hat{\mu_1} = \overline{x_1} = (5,8)'$$

Y la estimación de la matriz de varianzas, para lo que necesitamos:

$$S_0 = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 3 \end{pmatrix} \quad S_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

$$\hat{\Sigma} = \frac{(n_0-1)S_0 + (n_1-1)S_1}{n_0+n_1-2} = \frac{S_1+S_2}{2} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$$

Por último, la dirección proyección de la regla de fisher es

$$\omega = \hat{\Sigma^{-1}}(\overline{x}_1 - \overline{x}_0) = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

Entonces, utilizando la fórmula de clasificación de la regla de fisher, obtenemos:

$$2x_1 > (2,0) \begin{pmatrix} 4 \\ 7 \end{pmatrix} \rightarrow x_1 > 4$$

APARTADO B)

Como $x_1 = 2 \neg > 4$, el punto $x = (2,7)'$ lo clasificamos como P_0.

Observación: La frontera es una linea vertical. Las segundas coordenadas no importan nada, es curioso.

Ejercicio 4.2: *Considera los datos sobre enfermedades coronarias en Sudáfrica (infartos.RData). Calcula la función lineal discriminante de Fisher para clasificar entre sano (clase = 0) o enfermo (clase = 1) a un individuo en función de las 8 variables regresoras contenidas el fichero. Compara los coeficientes de las variables con los correspondientes a la regla de clasificación basada en regresión logística. ¿Son muy diferentes?*

```
# X <- matriz de los datos con $p = 8$ columnas y $n$ filas.
# clases <- vector de $0$ o $1$ si el individuo (columna) ha sufrido infarto
    (1) o no (0).
infartos <- lda(X,clases,prior=c(0.5,0.5))
```

Ejercicio 4.3: *Para 100 lirios, 50 de ellos correspondientes a la especie Versicolor (Y = 1) y otros 50 corres- pondientes a la especie Virginica (Y = 0) se ha medido la longitud (Long) y la anchura (Anch) del pétalo en milímetros. Con los datos resultantes se ha ajustado un modelo de regresión logística con el objetivo de clasificar en alguna de las dos especies un lirio cuya especie se desconoce a partir de las medidas de su pétalo. A continuación se muestra un resumen de los resultados (algunos valores han sido suprimidos o sustituidos por letras):*

```

glm(formula = y ~ Long + Anch, family = binomial)

Deviance Residuals:

Min      1Q      Median      3Q      Max
-1.8965923      -0.0227388      0.0001139      0.0474898      1.7375172

Coefficients:

     Estimate      Std.Error      z-value      Pr(>|z|)
(Intercept)      45.272      13.610      3.327      0.00088
Long      -5.755      2.306      *****      BBBB
Anch      -10.447      3.755      -2.782      0.00540


Null deviance: 138.629 on 99 degrees of freedom
Residual deviance: AAAA on 97 degrees of freedom

AIC: 26.564
```

a) *Escribe la fórmula de lo que en la salida de R se llama "Deviance residuals" y calcula la suma de estos residuos al cuadrado.*

b) *Calcula la desviación residual AAAA y contrasta, usando el método de razón de verosimilitudes, la hipótesis de que ningunad elas 2 medidas influte en la variable respuesta:* $H_0 : \beta_1 = \beta_2 = 0$

c) *Calcula el p-valor BBBB y contrasta a nivel* α *= 0,05 la hipótesis nula de que la longitud del pétalo no es significativa para explicar la respuesta.*

d) *Para un lirio se sabe que la longitud del pétalo es de 4.9 mm y la anchura es 1.5 mm. ¿En cuál de las dos especies se debe clasificar?*

APARTADO A)

Clacular la suma de los resudios y la desviación residual es lo mismo. Es el valor objetivo que sale al maximizar.

$$l(\hat{\beta}) = \sum D_i^2$$

Si recordamos la información de Akaike (3), tenemos:

$$AIC = -2l(\hat{\beta}) + 2(k+1) = -2A + 6 \to A = -2l(\hat{\beta}) = 26.564 - 6 = 20.564 = \sum_{i=1}^{n} D_i^2$$

APARTADO B)

$$138.629 - 20.564 = 118.065$$

Y comparamos este valor con $\chi^2_{2;0.05} = 5.99$, con lo que rechazamos la hipótesis y concluimos que las medidas de la planta influyen en la clase.

APARTADO C)

Vamos a utilizar el test de Wald (3.3) para contrastar $H_0 : \beta_1 = 0$

Tenemos $z = \frac{-5.755}{2.306} = -2.4957$.

$$B = P(|z| > 2.4957) \simeq 0.0128$$

APARTADO D)

Clasificar en $Y = 1$, entonces:

$$\frac{1}{1 + e^{-x'\beta}} > \frac{1}{2} \to \hat{\beta}x > 0$$

Es decir, la regla de clasificación logística en este caso es:

$$45.272 - 5.755 \cdot \underbrace{Long}_{4.9} - 10.447 \cdot \underbrace{Anch}_{1.5} > 0$$

Observación: ¿Y cuál es la probabilidad estimada de clasificar como $Y = 1$? No es lo mismo obtener en la regla anterior 0.001 o 0.9, que ambos son positivos. Para ello:

$$\frac{1}{1 + e^{-x'\beta}} = \frac{1}{1 + e^{1.4020}} = 0.19$$

Esto es un poco raro.

Ejercicio 4.4:

La dificultad de este problema radica en cómo introducir en R los datos para aplicar el comando glm.

Para ello, tenemos que meter n datos, por cada n insectos expuestos a un nivel de dosis.

Nuestro vector X entonces es:

$$X = \left(\underbrace{1.69, ..., 1.69}_{59}, \underbrace{1.7242, ..., 1.7242}_{60}, ... \right)$$

Y nuestro vector de clases sería:

$$Y = \left(\underbrace{1, ..., 1}_{6}, \underbrace{0, ..., 0}_{53}, ... \right)$$

En R sería:

```
y  = c(rep(1,6),rep(0,53),rep(1,13),rep(0,60-13),...)
```

Y ahora con los datos ya podemos calcular

```
reg <- glm(y ~ dosis, family='binomial')
```

Y ahora ya podemos utilizar:

$$\hat{P}(Y = 1|X = 1.8) = \frac{1}{1 + \exp -\hat{\beta}_0 - \hat{\beta}_1(1.8)} = 0.72$$

Ejercicio 4.5: *Para tratar la meningitis bacteriana es vital aplicar con urgencia un tratamiento con antibióticos. Por ello, es importante distinguir lo más rápidamente posible este tipo de meningitis de la menin- gitis vírica. Con el fin de resolver este problema se ajustó con R un modelo de regresión logística a las siguientes variables medidas en 164 pacientes del Duke University Medical Center:*

Nombre variable	*Descripción*
age	*Edad en años*
bloodgl	*Concentración de glucosa en la sangre*
gl	*Concentración de glucosa en el líquido cefalorraquídeo*
pr	*Concentración de proteína en el líquido cefalorraquídeo*
whites	*Leucocitos por mm 3 de líquido cefalorraquídeo*
polys	*Porcentaje de leucocitos que son leucocitos polimorfonucleares*
abm	*Tipo de meningitis: bacteriana (abm=1) o vírica (abm=0)*

El resultado del ajuste se muestra a continuación (algunos valores se han sustituido por letras):

a) *Calcula el valor de A en la salida anterior sabiendo que hay 68 pacientes con meningitis bacteriana en la muestra.*

b) *Calcula el valor de B en la salida anterior. A nivel $\alpha = 0.1$, ¿puede afirmarse que al aumentar la cantidad de leucocitos en el líquido cefalorraquídeo disminuye la probabilidad de que la meningitis sea de tipo vírico?*

c) *En un análisis realizado a un paciente de 15 años se han determinado los siguientes valores para el resto de variables:*

bloodgl	*119*
gl	*72*
pr	*53*
whites	*262*
polys	*41*

¿En cuál de los dos tipos de meningitis debe clasificarse este paciente?

Tenemos $k = 6$ y una proporción de $\frac{68}{164}$ individuos con meningitis bacteriana.

APARTADO A)

Null deviance $\equiv A \equiv D_0^2 \equiv -2\log(\hat{B}^{(0)})$ bajo $H_0 : \beta_1 = ... = \beta_k = 0$

Bajo H_0, $Y_1, .., Y_n \overset{iid}{\sim} B(1, p)$.

El E.M.V. de p es $\hat{p} = \frac{68}{164}$, entonces:

$$L(p) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} \to L(\hat{p}) = \sum_{i=1}^{n} \left[Y_i \log(\hat{p}) + (1 - Y_i)\log(1 - \hat{p})\right] = 68\log\left(\frac{68}{164}\right) + (164-68)\log\left(1 - \frac{}{1}\right.$$

APARTADO B)

B es el estadístico de Wald para la variable "white"

¿Cuál es nuestra H_0? Tenemos que si aumenta "white", entonces $P(Y = 0|x)$ disminuya. Esto no es la hipótesis nula, sino la hipótesis alternativa. Para construir la hipótesis nula, si "white" aumenta, etonces $H_0 : P(Y = 1|x)$ disminuya $\iff \beta_5 \leq 0$[2]

Entonces $B = z = \frac{\hat{\beta}_5}{e.t.(\hat{\beta}_5)} = \frac{0.00079971}{0.0005108} \simeq 1.56$

Para el contraste con $\alpha = 0.1$ y $H_0 : \beta_5 \leq 0$.

$$R = \{\}$$

APARTADO C)

$$\hat{\beta}_0 + \sum_{i=1}^{6} \hat{\beta}_i x_i = ... = -4.3136 < 0$$

Al ser negativo, lo clasificamos como vírico.

[2]Es importante darnos cuenta de que 0 es vírico y 1 bacteriano, al revés que la pregunta

Ejercicio 4.6:

Supongamos que la distribución de X condicionada a Y = 1 es normal con vector de medias μ 1 y matriz de covarianzas Σ, mientras que la distribución de X condicionada a Y = 0 es normal con vector de medias μ 0 y la misma matriz de covarianzas Σ (caso homocedástico). Demuestra que el error de la regla Bayes (error Bayes) del correspondiente problema de clasificación es:

donde $\Delta^2 = (\mu_0 - \mu_1)\Sigma^{-1}(\mu_0 - \mu_1)$ es el cuadrado de la distancia de Mahalanobis entre los dos vectores de medias y Φ es la función de distribución de una v.a. normal estándar. (Se supone que las probabilidades a priori de ambas poblaciones son iguales, π 0 = π 1 = 1/2)

En este caso, la regla de bayes es la regla de Fisher.

Definimos $g^* : \mathbb{R}^k \longmapsto \{0, 1\}$ definida como:

$$g^* = \begin{cases} 1 & \omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) > 0 \\ 0 & \omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) \leq 0 \end{cases} \quad \text{donde } \omega = \Sigma^{-1}(\mu_1 - \mu_0)$$

Para calcular el error, $L^* = P(g^*(x) \neq Y) = P(g^*(x) = 1, Y = 0) + P(g^*(x) = 0 Y = 1)$.

Vamos a calcular sólo uno de ellos:

$$P(g^*(x) = 1, Y = 0) = P(g^*(x) = 1 | Y = 0) \underbrace{P(Y = 0)}_{\frac{1}{2}}$$

Por otro lado,

$$P(g^*(x) = 1 | Y = 0) = P\left(\omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) > 0 | Y = 0 \right)$$

¿Y cuál es la distribución de $\omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) |_{Y=0}$? Es una **normal** (no se muy bien porqué)

Ahora, calculamos la media

$$\omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) = \frac{1}{2}(\mu_1 - \mu_0)' \Sigma^{-1}(\mu_0 - \mu_1) = ... = -\frac{1}{2}\Delta$$

y la varianza:

$$\mathbb{V}\left(\omega' x | Y = 0\right) = \omega' \Sigma \omega = ... = \Delta^2$$

Entonces,

$$\omega' \left(x - \frac{\mu_0 + \mu_1}{2} \right) |_{Y=0} \equiv N\left(-\frac{1}{2}\Delta, \Delta^2 \right)$$

Por último , siendo $z \sim N(0,1)$

$$P\left(\omega'\left(x - \frac{\mu_0+\mu_1}{2}\right) > 0 | Y = 0\right) = P\left(z > \frac{0 - \left(-\frac{\Delta^2}{2}\right)}{1}\right) = P\left(z > \frac{1}{2}\right) = 1 - \Phi\left(\frac{\Delta}{2}\right)$$

¿Tiene esto sentido? L^* es una función decreciente de $\Delta = \left((\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)\right)^{\frac{1}{2}}$. Esto quiere decir que si $\mu_0 = \mu_1$ (y como teníamos $\Sigma_1 = \Sigma_2 = \Sigma$), necesariamente $L^* = \frac{1}{2}$. [3]. Por otro lado, cuando $\Delta \to \infty$, tenemos un error que tiende a 0, consecuencia con sentido también.

[3] Si las distribuciones son exactamente iguales, no tenemos manera de distinguirlas

Apéndice B

Recordando

Esta sección ha sido obtenida de [Julián Moreno, 2013]

B.1. Estimador de máxima verosimilitud

En lo que sigue vamos a suponer que $\{X_n\}$ es una muestra formada por v.a.i.i.d. cuya distribución tiene una función de densidad o de masa $f(.;\theta_0)$ perteneciente a una familia de funciones $\{f(.;\theta) \diagup \theta \in \Theta\}$. θ_0 nos indica el valor real, y θ es un parámetro genérico.

Intuitivamente, lo que pensamos con este método es que la función de masa mide lo verosímil que es que salga un cierto parámetro.

Función de verosimilitud

Definición B.1.1 **Función de verosimilitud**. *También llamada likelihood function. Dada una muestra fija $\{x_n\}$, se define como*

$$L_n(\theta; x_1, \ldots, x_n) = L_n(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Estimador de máxima verosimilitud

Definición B.1.2 **Estimador de máxima verosimilitud**. *También llamado EMV o MLE (maximum likelihood estimator) es el argumento que maximiza la función de verosimilitud:*

$$\hat{\theta}_n = \hat{\theta}_n(x, \ldots, x_n) = \arg\max_{\theta \in \Theta} L_n(\theta; x_1, \ldots, x_n)$$

cuando ese máximo está bien definido.

Para evitar usar derivadas en un producto potencialmente muy largo, podemos maximizar el logaritmo de la verosimilitud, que es creciente y está bien definido porque la densidad es siempre mayor que cero, y los casos en los que sea cero no los estudiamos porque no ocurren (ocurren con probabilidad 0).

Apéndice C

Distribuciones, tablas

Tabla de la distribución Chi-cuadrado

g=grados de libertad p=área a la derecha
El valor x de la tabla cumple que para X es chi-cuadrado con g grados de libertad P(X>x)=p

	p										
g	**0.001**	**0.025**	**0.05**	**0.1**	**0.25**	**0.5**	**0.75**	**0.9**	**0.95**	**0.975**	**0.999**
1	10.827	5.024	3.841	2.706	1.323	0.455	0.102	0.016	0.004	0.001	0
2	13.815	7.378	5.991	4.605	2.773	1.386	0.575	0.211	0.103	0.051	0.002
3	16.266	9.348	7.815	6.251	4.108	2.366	1.213	0.584	0.352	0.216	0.024
4	18.466	11.143	9.488	7.779	5.385	3.357	1.923	1.064	0.711	0.484	0.091
5	20.515	12.832	11.07	9.236	6.626	4.351	2.675	1.61	1.145	0.831	0.21
6	22.457	14.449	12.592	10.645	7.841	5.348	3.455	2.204	1.635	1.237	0.381
7	24.321	16.013	14.067	12.017	9.037	6.346	4.255	2.833	2.167	1.69	0.599
8	26.124	17.535	15.507	13.362	10.219	7.344	5.071	3.49	2.733	2.18	0.857
9	27.877	19.023	16.919	14.684	11.389	8.343	5.899	4.168	3.325	2.7	1.152
10	29.588	20.483	18.307	15.987	12.549	9.342	6.737	4.865	3.94	3.247	1.479
11	31.264	21.92	19.675	17.275	13.701	10.341	7.584	5.578	4.575	3.816	1.834
12	32.909	23.337	21.026	18.549	14.845	11.34	8.438	6.304	5.226	4.404	2.214
13	34.527	24.736	22.362	19.812	15.984	12.34	9.299	7.041	5.892	5.009	2.617
14	36.124	26.119	23.685	21.064	17.117	13.339	10.165	7.79	6.571	5.629	3.041
15	37.698	27.488	24.996	22.307	18.245	14.339	11.037	8.547	7.261	6.262	3.483
16	39.252	28.845	26.296	23.542	19.369	15.338	11.912	9.312	7.962	6.908	3.942
17	40.791	30.191	27.587	24.769	20.489	16.338	12.792	10.085	8.672	7.564	4.416
18	42.312	31.526	28.869	25.989	21.605	17.338	13.675	10.865	9.39	8.231	4.905
19	43.819	32.852	30.144	27.204	22.718	18.338	14.562	11.651	10.117	8.907	5.407
20	45.314	34.17	31.41	28.412	23.828	19.337	15.452	12.443	10.851	9.591	5.921
21	46.796	35.479	32.671	29.615	24.935	20.337	16.344	13.24	11.591	10.283	6.447
22	48.268	36.781	33.924	30.813	26.039	21.337	17.24	14.041	12.338	10.982	6.983
23	49.728	38.076	35.172	32.007	27.141	22.337	18.137	14.848	13.091	11.689	7.529
24	51.179	39.364	36.415	33.196	28.241	23.337	19.037	15.659	13.848	12.401	8.085
25	52.619	40.646	37.652	34.382	29.339	24.337	19.939	16.473	14.611	13.12	8.649
26	54.051	41.923	38.885	35.563	30.435	25.336	20.843	17.292	15.379	13.844	9.222
27	55.475	43.195	40.113	36.741	31.528	26.336	21.749	18.114	16.151	14.573	9.803
28	56.892	44.461	41.337	37.916	32.62	27.336	22.657	18.939	16.928	15.308	10.391
29	58.301	45.722	42.557	39.087	33.711	28.336	23.567	19.768	17.708	16.047	10.986
30	59.702	46.979	43.773	40.256	34.8	29.336	24.478	20.599	18.493	16.791	11.588
35	66.619	53.203	49.802	46.059	40.223	34.336	29.054	24.797	22.465	20.569	14.688
40	73.403	59.342	55.758	51.805	45.616	39.335	33.66	29.051	26.509	24.433	17.917
45	80.078	65.41	61.656	57.505	50.985	44.335	38.291	33.35	30.612	28.366	21.251
50	86.66	71.42	67.505	63.167	56.334	49.335	42.942	37.689	34.764	32.357	24.674
55	93.167	77.38	73.311	68.796	61.665	54.335	47.61	42.06	38.958	36.398	28.173
60	99.608	83.298	79.082	74.397	66.981	59.335	52.294	46.459	43.188	40.482	31.738
65	105.988	89.177	84.821	79.973	72.285	64.335	56.99	50.883	47.45	44.603	35.362
70	112.317	95.023	90.531	85.527	77.577	69.334	61.698	55.329	51.739	48.758	39.036
75	118.599	100.839	96.217	91.061	82.858	74.334	66.417	59.795	56.054	52.942	42.757
80	124.839	106.629	101.879	96.578	88.13	79.334	71.145	64.278	60.391	57.153	46.52
85	131.043	112.393	107.522	102.079	93.394	84.334	75.881	68.777	64.749	61.389	50.32
90	137.208	118.136	113.145	107.565	98.65	89.334	80.625	73.291	69.126	65.647	54.156
95	143.343	123.858	118.752	113.038	103.899	94.334	85.376	77.818	73.52	69.925	58.022
100	149.449	129.561	124.342	118.498	109.141	99.334	90.133	82.358	77.929	74.222	61.918

Apéndice D

Prácticas

Se incluyen las soluciones de las prácticas:

Práctica 1 Estadística II

Alberto Parramón Castillo

Introducimos en una variable los datos de la tabla Iris. Sólo las 50 primeras filas, menos la quinta columna: longitud del sépalo - anchura del sépalo - longitud del pétalo - anchura del pétalo

```
datos <- iris[1:50,-5]
head(datos)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

Ejercicio 1

Calcula el vector de medias muestral y las matrices de covarianzas y de correlaciones (cor) muestrales. ¿Entre qué par de variables es más alta la correlación? ¿Qué variable tiene la mayor varianza?

A) Vector de medias:

```
mediasIris <- colMeans(datos)
mediasIris
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##        5.006        3.428        1.462        0.246
```

B) Matriz de covarianzas:

```
covIris <- cov(datos)
covIris
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.12424898 0.099216327  0.016355102 0.010330612
## Sepal.Width    0.09921633 0.143689796  0.011697959 0.009297959
## Petal.Length   0.01635510 0.011697959  0.030159184 0.006069388
## Petal.Width    0.01033061 0.009297959  0.006069388 0.011106122
```

C) Matriz de correlaciones:

```
corIris <- cor(datos)
corIris
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

D) ¿Entre qué par de variables es más alta la correlación?

Entre longitud de sepalos y anchura de sepalos: 0.7425467

E) ¿Qué variable tiene la mayor varianza?

La anchura de sepalos

Ejercicio 2

Calcula las distancias de Mahalanobis entre cada uno de los lirios y el vector de medias. Representa los datos, usando el color rojo para el 25 % de los lirios más lejanos al vector de medias.

A) Utilizamos la función de Mahalanobis con parámetros: los datos, el vector de medias, y la matriz de covarianzas:

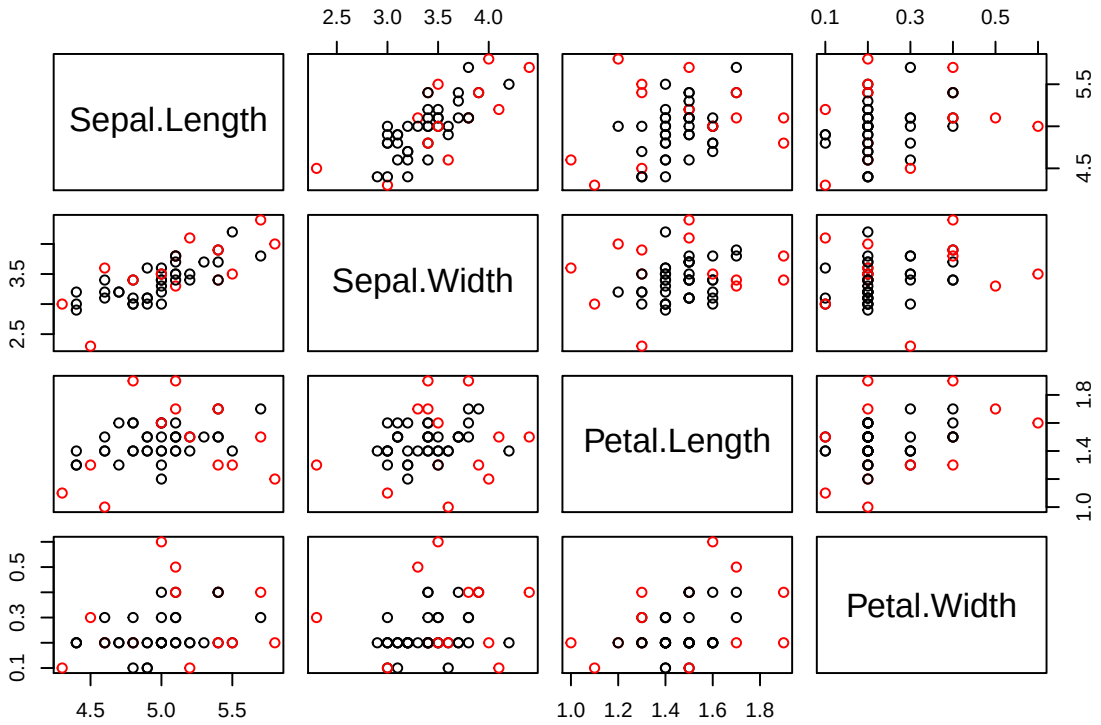
```
distancias <- mahalanobis(datos, mediasIris, covIris)
```

B) Utilizamos la función *summary*, que nos devuelve un vector cuyo quinto elemento es el tercer cuartil de los datos que le hayas pasado por argumento, en este caso las distancias.

```
cuartil3 <- summary(distancias)[5]
```

Creamos el vector de colores y pintamos con plot:

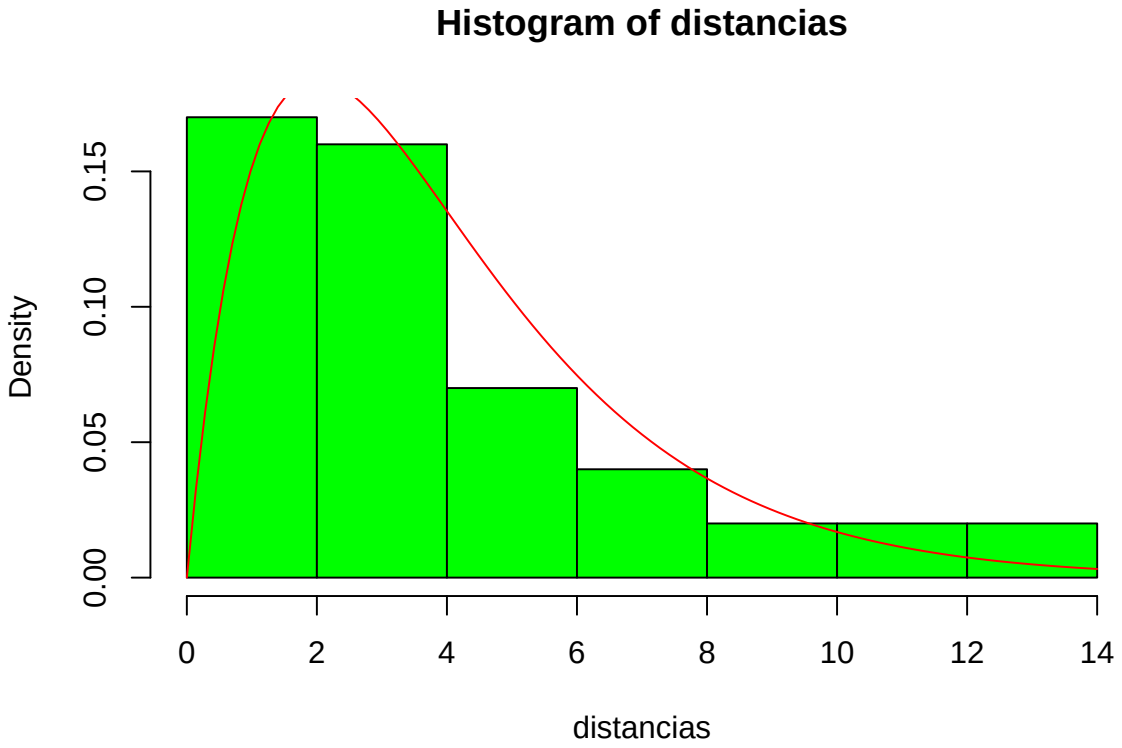
```
colores <- vector('character', length=50)
for(i in 1:50){
  if(distancias[i]>cuartil3){
    colores[i] <- 'red'
  }else{
    colores[i] <- 'black'
  }
}
pairs(datos, col=colores)
```



Ejercicio 3

Representa un histograma de las distancias y compáralo con la función de densidad de una variable χ^2 con 4 grados de libertad.

```
hist(distancias, col = "green", breaks = 8, freq=FALSE)
curve( dchisq(x, df=4), col='red', add=TRUE)
```



Ejercicio 4

Genera 100 observaciones con distribución normal bidimensional con vector de medias el origen y matriz de covarianzas:

$$\Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$$

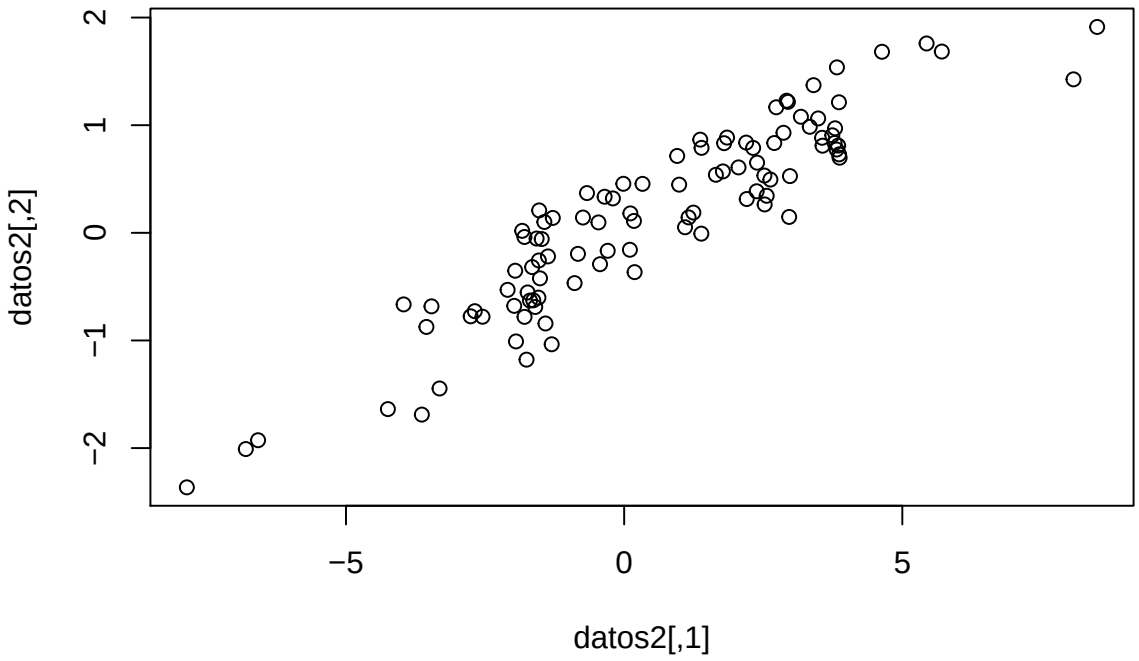
Representa la nube de puntos generados, su vector de medias y su matriz de covarianzas.

A) Obtenemos las 100 observaciones a partir de los datos del enunciado, siendo *mu* el vector de medias, *sigma* la matriz de covarianzas y n el número de observaciones:

```
set.seed(9111) #Esto establece una semilla para que siempre salgan los mismos datos aleatorios
library(MASS) #paquete necesario
n <-100
mu <- c(0,0)
sigma <-matrix(c(10,3,3,1),2)
datos2 <- mvrnorm(n,mu,sigma)
```

Representamos la nube de puntos:

```
plot(datos2)
```



B) Calculamos y representamos su vector de medias obtenido con los datos generados

```
medias = colMeans(datos2)
medias
```

```
## [1] 0.5300716 0.1524980
```

C) Calculamos y representamos la matriz de covarianza obtenida con los datos generados

```
covarianza = cov(datos2)
covarianza
```

```
##          [,1]      [,2]
## [1,] 8.669332 2.3585971
## [2,] 2.358597 0.7536912
```

Ejercicio 5

Para la misma distribución del apartado anterior, calcula el valor esperado teórico de la segunda coordenada respecto de la primera. Si no lo conocieras y solo dispusieras de los datos generados. ¿Cómo lo estimarías? Calcula el valor resultante para el estimador que has propuesto.

Si suponemos que queremos calcular el valor esperado de $X_2|X_1$. Utilizaremos las siguientes fórmulas generales.

$$\mu_{2.1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$$

$$\Sigma_{2.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

A) Valor esperado teórico para $X_2|X_1$, tenemos el vector de medias y la matriz de covarianzas siguiente:

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 1 \end{pmatrix}$$

Obtenemos:

$$\mu_{2.1} = 0 + \frac{3}{10}(X_1)$$
$$\Sigma_{2.1} = 1 - \frac{3}{10}3 = \frac{1}{10}$$

B) Valor esperado estimado a partir de las observaciones para $X_2|X_1$, tenemos el vector de medias y la matriz de covarianzas siguiente:

$$\mu = \begin{pmatrix} 0.53 \\ 0.15 \end{pmatrix}, \Sigma = \begin{pmatrix} 8.66 & 2.35 \\ 2.35 & 0.75 \end{pmatrix}$$

Obtenemos:

$$\mu_{2.1} = 0.15 + \frac{2.35}{8.66}(X_1 - 0.53) = 0.006 + 0.27X_1$$
$$\Sigma_{2.1} = 0.75 - \frac{2.35}{8.66}2.35 = 0.11$$

Práctica 2 Estadística II

Alberto Parramón Castillo

En primer lugar cargamos los datos en la variable *goles*0809

```
load('goles0809.RData')
```

Contrastes basados en la distribución χ^2

Ejercicio 1

Contrasta si la diferencia de goles entre los dos equipos que juegan cada partido sigue una distribución uniforme.

Así tenemos como hipótesis nula: $H_0 : X \sim Uniforme$.

Guardamos los goles en casa y los goles fuera de casa en variables diferentes. Los restamos y sacamos su valor absoluto (ya que lo que nos importa es la diferencia y no el el signo), después clasificamos esos goles en una tabla:

```
golesCasa <- goles0809$casa
golesFuera <- goles0809$fuera
difGoles <- golesCasa - golesFuera
difGoles <- abs(difGoles)
difGoles <- table(difGoles)
difGoles
```

```
## difGoles
##   0   1   2   3   4   5   6
##  83 160  78  38  13   5   3
```

Agrupamos las dos ultimas columnas en una sola:

```
difGoles <- c(difGoles[1:5], sum(difGoles[6:7]))
names(difGoles)[6] <- '>4'
difGoles
```

```
##   0   1   2   3   4  >4
##  83 160  78  38  13   8
```

Por defecto la función chisq.test te calcula la diferencia de goles suponiendo una distribución uniforme :

```
chisq.test(difGoles)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  difGoles
## X-squared = 255.53, df = 5, p-value < 2.2e-16
```

Sale un p-valor muy cercano a 0, por tanto para casi cualquier nivel de significación α se rechaza la hipótesis nula. Rechazamos la idea de que la diferencia de goles siga una distribución uniforme.

Ejercicio 2

Contrasta si la diferencia de goles entre los dos equipos que juegan cada partido sigue una distribución de Poisson.

Así tenemos como hipótesis nula: $H_0 : X \sim Poisson(lambda)$.

Al igual que antes sacamos la tabla de los goles:

```
difGoles <- golesCasa - golesFuera
difGoles <- abs(difGoles)
difGoles <- table(difGoles)
```

Ahora calculamos el EMV de λ:

```
clases = seq(0,6)
n = sum(difGoles)
lambda = sum(clases*difGoles)/n
lambda
```

```
## [1] 1.381579
```

Calculamos las probabilidades estimadas de cada clase, así como las esperanzas estimadas de cada clase:

```
prob = dpois(clases, lambda)
esp = n*prob
esp
```

```
## [1]  95.449022 131.870360  91.094656  41.951486  14.489823   4.003767
## [7]   0.921920
```

Agrupamos las clases 6 y 7 ya que valen menos de 5.

```
difGoles <- c(difGoles[1:5], sum(difGoles[6:7]))
prob <- c(prob[1:5], 1-sum(prob[1:5]))
esp <- c(esp[1:5], n-sum(esp[1:5]))
```

Obtenemos el estadístico y el p-valor, pero el p-valor que obtiene R en las hipótesis nulas compuestas no es correcto. Por ello lo calculamos con la tabla de la χ^2 con $k-1-r$ grados de libertad. Como k=6 (que son las clases) y r=1 (que es la dimensión del EMV), nos queda 4:

```
t=chisq.test(difGoles, p=prob)$statistic
pvalor = 1-pchisq(t,4)
pvalor
```

```
##  X-squared
## 0.02044257
```

El p-valor es 0.02, por tanto, a veces rechazaríamos la hipótesis nula, es decir, rechazaríamos que los datos siguen una distribución de Poisson, y otras veces no. Dependerá del nivel de significación que queramos asumir, para niveles de significacion $\alpha > 0.02$ rechazaríamos la hipótesis nula.

Por ejemplo, si tenemos un nivel de significación $\alpha = 0.01$, no rechazaríamos la hipótesis nula, ya que $\alpha = 0.01$ quiere decir que queremos rechazar la hipótesis nula con una probabilidad máxima de equivocarnos del 1%, sin embargo, el análisis que hemos obtenido, nos da un $p - valor = 0.02$, eso quiere decir, que al menos tenemos que afrontar una probabilidad de equivocarnos al rechazar la hipótesis nula de un 2%.

Con nivel de significación $\alpha = 0.05$ si rechazaríamos la hipótesis nula, ya que asumimos una probabilidad máxima de equivocarnos del 5% y el p-valor nos dice que tenemos solo un 2% de probabilidades de equivocarnos.

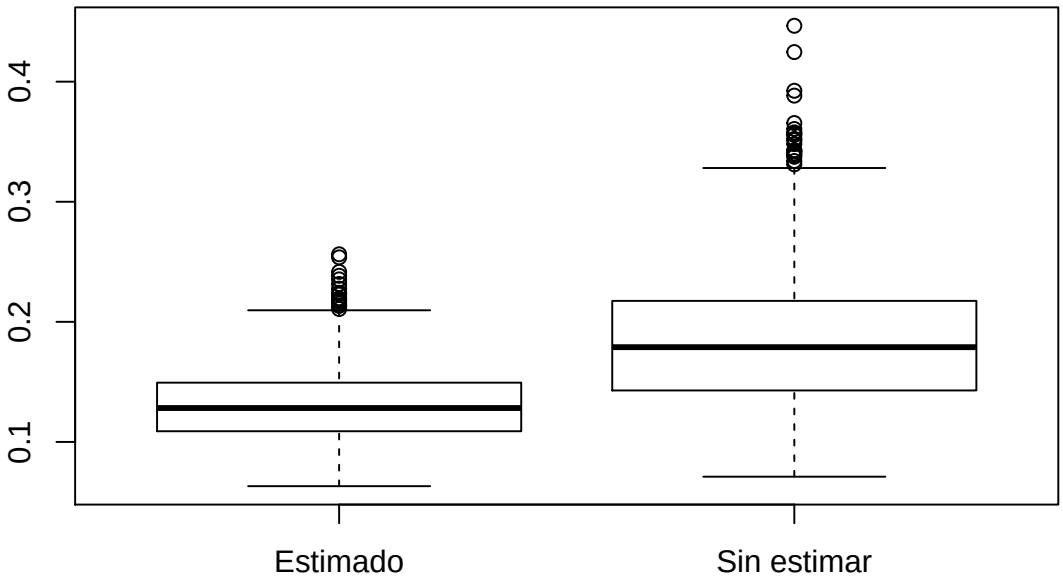
Contraste de Kolmogorov-Smirnov

```
ksnoest <- function(datos){
  y <- ks.test(datos,pnorm)$statistic
  return(y)
}

ksest <- function(datos){
  mu <- mean(datos)
  stdev <- sd(datos)
  y <- ks.test(datos, pnorm, mean=mu, sd=stdev)$statistic
  return(y)
}

B <- 1000
n <- 20
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest) #El 2 es para hacerlo por columnas
tnoest <- apply(datos, 2, ksnoest)

boxplot(test, tnoest, names=c("Estimado", "Sin estimar"))
```



Ejercicio 1

Claramente las distribuciones de test y de tnoest son diferentes, por lo que no podemos usar las mismas tablas para hacer el contraste en las dos situaciones. ¿En cuál de los dos casos se obtienen en media valores menores? ¿Podrías dar una razón intuitiva?

Lo que representamos en las cajas es el valor del estadístico. De media se obtienen valores más pequeños en el estimado.

Sabemos que un valor del estadístico pequeño implica un p-valor grande, y un p-valor grande implica que la probabilidad de equivocarnos si decidimos rechazar la hipótesis nula es grande. Por otro lado, sabemos que el p-valor (y por tanto el valor del estadístico) dependen de los datos de partida y de la hipótesis nula. En este caso, los datos de partida son los mismos en ambos casos (ambos proceden de generaciones aleatorias de muestras de una $N(0,1)$) por tanto la diferencia entre la primera caja y la segunda esta relacionada con la hipótesis nula (H_0):

En la segunda caja (Sin estimar), simplemente comparamos los datos aleatorios que generamos con una distribución $N(0,1)$, por tanto, los datos pueden haber salido un poco diferentes a esa $N(0,1)$, ya que son aleatorios; pero como provienen precisamente de una $N(0,1)$ es de esperar que se parezcan bastante a esta y que el valor del estadístico sea bastante grande, y que el p-valor sea por tanto bastante pequeño.

Mientras que en la primera caja (Estimado) estimamos la media y la desviación típica de los datos, y después suponemos como H_0 que los datos siguen una distribución normal de media y desviación típica las estimadas a partir de los datos. Por tanto, es natural que los datos se parezcan mucho más a esa distribución dada por H_0 que los de la segunda caja, y por tanto, es bastante intuitivo pensar que el p-valor va a salir bastante grande, y por tanto el valor del estadístico bastante pequeño.

En ambos casos, seguramente no rechazaríamos la hipótesis nula para valores de α habituales (0.01 o 0.05).

Ejercicio 2

Imagina que estimamos los parámetros y usamos las tablas de la distribución del estadístico de Kolmogorov-Smirnov para hacer el contraste a nivel α. El verdadero nivel de significación, ¿es mayor o menor que α?

En la caja de los estimados tenemos los valores de los estadísticos más pequeños que en la caja de los contrastes sin estimar. Por tanto, los p-valores son más altos en los contrastes estimados que en los que están sin estimar. Por tanto, para un α en los estimados, el α que salga en los que están sin estimar será más pequeño.

Esto es intuitivo si volvemos a interpretar α como la probabilidad máxima que queremos asumir de equivocarnos al rechazar la hipótesis nula. Como hemos visto antes, la hipótesis nula es menos rechazable en de los contrastes con los datos estimados que en los que están sin estimar. Por tanto, escojo un valor α en los estimados, que representará una probabilidad máxima de equivocarme al rechazar H_0 de un $x\%$. Este α escogido llevará asociado un nivel cítico (valor en la tabla) en el contraste de los datos estimados, y ese mismo nivel crítico en el contraste de los parámetros sin estimar llevará asociado un valor α más pequeño que el anterior. Esto es razonable ya que en el contraste sin estimar la probabilidad de equivocarnos al rechazar H_0 es algo menor.

Ejercicio 3

Para resolver el problema se ha estudiado la distribución en el caso de muestras normales con parámetros estimados. Es lo que se conoce como contraste de normalidad de Kolmogorov-Smirnov-Lilliefors (KSL) (véase, por ejemplo, Peña (2001), pag. 471 y Tabla 9). Según la tabla del estadístico KSL, el nivel crítico para $\alpha = 0.05$ y n = 20 es 0.190. Esto significa que el porcentaje de valores test} mayores que 0.19 en nuestra simulación debe ser aproximadamente del 5%. Compruébalo haciendo sum(test > 0.19)/B}. Haz una pequeña

simulación similar a la anterior para aproximar el nivel de significación del contraste KSL cuando se utiliza un valor crítico 0.12 para muestras de tamaño 40.

Si asumimos un $\alpha = 0.05$ es que asumimos una probabilidad máxima de equivocarnos al rechazar H_0 del 5%. Vamos a contrastar datos que provienen de una distribución normal, con la hipótesis nula de que siguen una distribución normal de parámetros μ y *sd* estimados empirícamente. Por tanto, si rechazamos H_0 claramente nos estamos equivocando, y la prbabilidad de equivocarnos al rechazar H_0 viene determinada por α. Por tanto, si $\alpha = 0.05$, lleva asociado un nivel crítico de 0.19, quiere decir que sólo nos vamos a encontrar con datos que provoquen un valor estadístico T>0.19 (es decir, entrando en la región de rechazo) en un 5% de los casos que estudiemos.

Lo comprobamos:

```
B <- 1000
n <- 20
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest)
sum(test>0.19)/B
```

```
## [1] 0.056
```

Ahora vamos a calcular α sabiendo que el nivel crítico es 0.12 y las muestras son de tamaño n=40:

```
B <- 1000
n <- 40
datos <- matrix(rnorm(n*B), n)
test <- apply(datos, 2, ksest)
alpha = sum(test>0.12)/B
#Mostramos el valor de alpha:
alpha
```

```
## [1] 0.135
```

Ejercicio 4

Genera B = 10000 muestras de tamaño n = 30 de una distribución exponencial de media 1 y utilízalas para determinar en este caso la potencia aproximada del test de Kolmogorov-Smirnov con $\alpha = 0.05$ para $H_0 \equiv N(1, 1)$. El comando rexp() puede utilizarse para generar los datos exponenciales).

Obtenemos de la tabla de Kolmogorov-Smirnov el valor para $\alpha = 0.05$: $D_{\alpha=0.05} = 0.24$.

Comprobamos que 0.24 es el nivel crítico para $\alpha = 0.05$, para ello, generamos muestras de una N(1,1) y comprobamos que la probabilidad de rechazar $H : 0$ siendo esta verdadera es de un 5%:

```
ksej4_1 <- function(datos){
  y <- ks.test(datos, pnorm, mean=1, sd=1)$statistic
  return(y)
}
```

```
B <- 10000
n <- 30
datos <- matrix(rnorm(n*B, mean=1, sd=1), n)
test <- apply(datos, 2, ksej4_1)
sum(test>0.24)/B
```

```
## [1] 0.0483
```

Vemos que nos sale aproximadamente un 5%. Ahora vamos con los que nos pide el enunciado. La potencia del contraste es ver cuántas veces se rechaza la hipótesis nula:

```
ksej4_2 <- function(datos){
  y <- ks.test(datos, pnorm, mean=1, sd=1)$statistic
  return(y)
}
```

```
B <- 10000
n <- 30
datos <- matrix(rexp(n*B), n)
test <- apply(datos, 2, ksej4_2)
sum(test>0.24)/B
```

```
## [1] 0.2886
```

Por tanto tenemos una potencia del contraste de aproximadamente un 29%

Hoja 2 de ejercicios

Ejercicio 9

A finales del siglo XIX el físico norteamericano Newbold descubrió que la proporción de datos que empiezan por una cifra d, p(d), en listas de datos correspondientes a muchos fenómenos naturales y demográficos es aproximadamente:

$$p(d) = log_{10}\left(\frac{d+1}{d}\right), (d = 1, 2, ..., 9)$$

Por ejemplo, $p(1) = log_{10}2 \approx 0,301030$ es la frecuencia relativa de datos que empiezan por 1. A raíz de un artículo publicado en 1938 por Benford, la fórmula anterior se conoce como ley de Benford. El fichero poblacion.RData incluye un fichero llamado poblaciones con la población total de los municipios españoles, así como su población de hombres y de mujeres. (Indicación: Puedes utilizar, si te sirven de ayuda, las funciones del fichero benford.R).

Aquí tenemos las funciones del fichero benford.R

```
#-----------------------------------------------------
#
# Una funcion para contar las frecuencias:
# Dado un vector x, esta funcion calcula la frecuencia de valores
# que empiezan por 1, 2, ..., 9
#
#-----------------------------------------------------
benford = function(x){
    n = length(x)
    proporcion = numeric(9)
    for (i in 1:9){
        proporcion[i] = sum(substr(x,1,1)==as.character(i))
    }
```

```
    return(proporcion)
}

#-----------------------------------------------------------------
# Una funcion para contar las frecuencias de los dos primeros digitos
# Dado un vector x, esta funcion calcula la tabla de frecuencias de los valores
# de los pares (i,j) donde i = 1, 2, ..., 9 y j = 0, 1, ..., 9
# (solo considera valores mayores o iguales que 10)
#
#-----------------------------------------------------------------
benford2 = function(x){
    x = x[x>=10]
    n = length(x)
    proporcion = matrix(0,9,10)
    digitos = substr(x,1,2)

    for (i in 1:9){
        for (j in 1:10){
            proporcion[i,j] = sum(digitos==paste(i,j-1,sep=''))/n
        }
    }
    colnames(proporcion) = paste(0:9)
    rownames(proporcion) = paste(1:9)
    return(proporcion)
}
```

En primer lugar cargamos el fichero benford.R

```
load('poblacion.RData')
```

A) Contrasta a nivel $\alpha = 0{,}05$ la hipótesis nula de que la población total se ajusta a la ley de Benford.

Definimos una función que nos devuelve las probabilidades de cada clase (dígito) según H_0, es decir, suponiendo que los dígitos siguen la distribución dada por Benford:

```
probBenford = function(){
    proporcion = numeric(9)
    for (i in 1:9){
        proporcion[i] = log10((i+1)/i)
    }
    return(proporcion)
}
```

Utilizamos el contraste de bondad de ajuste basados en la distribución χ^2.

```
pobTotalFrecuencias <- benford(poblaciones$pobtotal)
prob = probBenford()
chisq.test(pobTotalFrecuencias, p=prob)
```

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  pobTotalFrecuencias
## X-squared = 13.5, df = 8, p-value = 0.09575
```

Como el p-valor es 0.095, que es mayor que 0.05, no podemos rechazar la hipótesis nula H_0 a nivel de significacion $\alpha = 0.05$.

B) Repite el ejercicio pero considerando sólo los municipios de más de 1000 habitantes.

```
pob1000 = poblaciones$pobtotal[poblaciones$pobtotal > 1000]
pob1000Frecuencias <- benford(pob1000)
prob = probBenford()
chisq.test(pob1000Frecuencias, p=prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  pob1000Frecuencias
## X-squared = 298.91, df = 8, p-value < 2.2e-16
```

Como el p-valor es 2.2e-16, que es menor que 0.05, podemos rechazar la hipótesis nula H_0 a nivel de significacion $\alpha = 0.05$.

C) Considera las poblaciones totales (de los municipios con 10 o más habitantes) y contrasta a nivel $\alpha =$ 0,05 la hipótesis nula de que el primer dígito es independiente del segundo.

```
n = length(poblaciones$pobtotal[poblaciones$pobtotal >= 10])
frecuencias = n*benford2(poblaciones$pobtotal)
chisq.test(frecuencias)
```

```
##
##  Pearson's Chi-squared test
##
## data:  frecuencias
## X-squared = 120.52, df = 72, p-value = 0.0002974
```

Como el p-valor es 0.0002974, que es menor que 0.05, podemos rechazar la hipótesis nula H_0 a nivel de significacion $\alpha = 0.05$.

Bibliografía

Guillermo Julián Moreno. Apuntes Estadística I. http://github.com/Vicdejuan/Apuntes, 2013. Apuntes UAM.

Índice alfabético